University of New Mexico
UNM Digital Repository

Mathematics & Statistics ETDs | Electronic Theses and Dissertations

Spring 5-11-2020

A Statistical Analysis of the UNM FACETS Design Identity & Beliefs Survey Data

Clarissa A. Sorensen-Unruh
University of New Mexico - Main Campus

Follow this and additional works at: https://digitalrepository.unm.edu/math_etds

 Part of the Applied Mathematics Commons, Educational Assessment, Evaluation, and Research Commons, Engineering Education Commons, Higher Education Commons, Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation

Sorensen-Unruh, Clarissa A.. "A Statistical Analysis of the UNM FACETS Design Identity & Beliefs Survey Data." (2020). https://digitalrepository.unm.edu/math_etds/149

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at UNM Digital Repository. It has been accepted for inclusion in Mathematics & Statistics ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.

Clarissa Sorensen-Unruh
Candidate

Mathematics and Statistics
Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by Thesis Committee:

James Degnan, *Chairperson*

Erik Erhardt

Vanessa Svihla

A Statistical Analysis of the UNM FACETS Design Identity & Beliefs Survey Data

by

Clarissa A. Sorensen-Unruh

Bachelor of Science, Biochemistry, Trinity University, 1998
Master of Science, Chemistry, University of New Mexico, 2000

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science in Statistics (MS-STAT)

The University of New Mexico
Albuquerque, New Mexico

May 2020

This manuscript is lovingly
dedicated to
my wife and my son –
my favorite people –
who
continually and gently
encourage me to be
my best self.

Acknowledgements

There are too many people who have supported me through the last two years of doing the work for this thesis to acknowledge on this one page. Most of you know who you are, so if your name is omitted, it is not from lack of appreciation but a lack of time and space.

I acknowledge my in-town family – my wife and son, who have heard so much about statistics in the last two years and love me nonetheless; my mother, who read this thesis four or more times as my best editor and, with the right combination of mom + editor, catalyzed my writing of the thesis when I didn't want to write it; my sister, who acts interested and supports infinitely; and my father, who is always willing to celebrate my achievements.

I acknowledge my friend support group – Heather, Phil, Anna, Ivy, MaryKay, Sushilla, Steve, Maha, etc. – who have gently asked about the thesis since I began the work on it.

Thank you to my friends and family for loving and supporting me throughout this entire process.

I acknowledge my M.S. advisor, Dr. James Degnan, who was compassionate and kind at all the times I needed it most.

I acknowledge my OILS Ph.D. advisor, Dr. Vanessa Svihla, who allowed me to use the survey data she had collected and journeyed with me through my analysis. I truly appreciate your help and support, Vanessa.

I acknowledge my last Statistics committee member, Dr. Erik Erhardt, who taught me R and with whom I began my journey of how to think statistically in ADA I and II.

Thank you, Statistics Committee Members.

A Statistical Analysis of the UNM FACETS Design Identity & Beliefs Survey Data

Clarissa Sorensen-Unruh
Bachelor of Science, Biochemistry
Master of Science, Chemistry
Master of Science, Statistics

Abstract

The NSF-funded FACETS (Formation of Accomplished Chemical Engineers for Transforming Society, NSF Award 1623105) grant aims to transform the undergraduate engineering experience in the Department of Chemical and Biological Engineering at the University of New Mexico to address attrition within engineering majors, especially among underserved populations (Brainard & Carlin, 1998). The UNM FACETS Design Identity & Beliefs survey, an assessment tool used as part of the research of the grant, generated the dataset used in this study. I performed several different statistical analyses on the dataset, including confirmatory factor analysis (CFA), principal component analysis (PCA), and cluster analysis. The information obtained from these analyses was used to shorten the survey by eliminating ten questions that did not cluster with other questions asking about the same construct. Regression analysis and ANOVA techniques were used to build a model to predict student persistence using both the longer and the shortened survey.

Table of Contents

List of Figures

List of Tables

Introduction

Attrition within undergraduate engineering majors is a well-studied problem (Geisinger & Raman, 2013; Santiago & Hensel, 2012; Bernold, Spurlin, & Anson, 2007; Huang, Taddese, Walter, & Peng, 2000). The NSF-funded FACETS (Formation of Accomplished Chemical Engineers for Transforming Society, NSF Award number 1623105) grant aims to transform the undergraduate engineering experience in the Department of Chemical and Biological Engineering at the University of New Mexico to address this attrition, especially among underserved populations (Brainard & Carlin, 1998; Huang, Taddese, Walter, & Peng, 2000). The FACETS grant has three major components*:*

> (1) introducing "CIRE" (Community-, Industry-, Research-, and/or Entrepreneurship-based) design challenges in the core curriculum to increase community engagement and 'engage students in developing their sociotechnical awareness and attract diverse, native and rural populations into engineering' (2) conducting professional development institutes that will train faculty and graduate students; workshops will be led by experts from industry and national laboratories, from the learning sciences, from engineering education and multicultural studies, and (3) creating a digital badging that will help students take ownership of their competencies and develop engineering identities. (Datye, Chi, Han, Svihla, & Kang, 2016)

This study analyzes a survey intended to track impacts of these changes as part of the UNM FACETS grant. The survey that generated the dataset asks students about their engineering identities by conceptualizing different facets of their crystallized identity. Tracy and Trethewey (2005, p. 189) characterize crystallized identity as multifaceted, "ongoing, emergent and not entirely predictable." They reframe the dichotomy of real and fake selves in terms of every aspect of identity. The use of the crystallized identity concept within this survey was based in a larger

overarching hypothesis guiding many projects under the FACETS grant: the idea that if students, particularly those from marginalized and underserved groups, could identify with what engineers do (through the CIRE design challenges), then those students would persist in the engineering degree.

The survey, which will be referred to as The Survey or as the Design Identity & Beliefs survey (Study Design, n.d.) within this paper, included several constructs the PIs (Principal Investigators) of the FACETS grant wanted to measure, including: 1. Knowledge of/beliefs about design practices; 2. Engineering design self-efficacy; 3. Intent to persist in engineering; 4. Social factors related to persistence; 5. Degree choice; and 6. Professional engineering identity. The survey items were taken from previous studies, including Mosborg et. al (2005), Carberry et. al (2010), Nocito-Gobel et. al (2005), Pierrakos et. al (2009), and Sheppard et. al (2010). Mosborg, Adams, Kim, Atman, Turns, and Cardella (2005) looked at how expert engineers interpret and use a block diagram, which is a flowchart for solving engineering problems emphasized in many textbooks. In this qualitative study, the expert designers were asked to rate twenty-seven design statements that helped describe definitions of design by using a five-point Likert scale (1 = strongly disagree; 5 = strongly agree). These design statements emphasized construct 1 - knowledge of/beliefs about design practices in engineering.

Carberry, Lee, and Ohland (2010) developed and validated a survey that measured engineering design self-efficacy (construct 2). The thirty-six question self-efficacy survey was administered to 202 engineering students via an online survey tool and was validated through content, criterion-related, and construct validity

measures. Throughout their study, Carberry et. al were able to show that the self-efficacy instrument they designed had a high amount of reliability in measurement and high validity in all three criteria.

The engineering faculty at the University of New Haven redesigned a first-year engineering course into a project-based course. Nocito-Gobel, Collura, Daniels, and Orabi (2005) surveyed students in both the project-based course (the intervention) and traditional delivery (the control) based on their perceptions of the engineering profession and their engineering field preparation in a pre-/post-test format. The survey used was a modification of the Pittsburgh Freshman Engineering Attitude Instrument, a validated instrument used since 1993. This study emphasized constructs 3, 4, 5, and 6, or intent to persist in engineering, social factors related to persistence, degree choice, and professional engineering identity, respectively.

Pierrakos, Beam, Constantz, Johri, and Anderson (2009) conducted forty-five interview and focus groups with both STEM and non-STEM freshman students at a large, rural university. The qualitative study showed that those students who typically persisted in engineering majors had had more engineering-related experiences, and therefore knew more about the engineering profession overall. Those students who didn't persist often had misperceptions as to what engineers actually do. Many of the student comments shared within this paper became items on the Design Identity & Beliefs survey (Study Design, n.d.) developed by the FACETS PIs. Constructs developed using this study include 4, 5, and 6, or social

factors related to persistence, degree choice, and professional engineering identity, respectively.

Sheppard, Gilmartin, Chen, Donaldson, Lichtenstein, Eris, Lande, and Toye (2010) analyzed the implementation of a massive survey instrument called APPLES (Academic Pathways of People Learning Engineering Survey), which probes the following constructs: confidence and perceived importance of certain fundamental skills, motivation, academic and professional persistence, and professional engineering identity knowledge. These constructs were chosen to better understand undergraduate students' experience within the engineering curricula and major. The survey instrument was further validated from the previous instruments (APPLES1 and PEI (or the Persistence in Engineering survey)). Twenty-one institutions participated in the survey. The UNM FACETS Design Identity & Beliefs survey constructs developed from this survey and report include: 3 - intent to persist in engineering; 4 - social factors related to persistence; 5 - degree choice; and 6 - professional engineering identity

The central problem this study addresses is that in the absence of a single comprehensive survey that has been subject to validation procedures, the FACETS study PIs drew from multiple extant surveys (of varied quality) to measure constructs reportedly salient to the problem. This resulted in a very long survey, and the length of the survey may contribute to a collection of data that is not representative of the student population due to nonresponsiveness and survey fatigue.

Survey fatigue has been studied extensively within the literature and takes many forms. Participants can experience survey fatigue by: 1. filling out a survey that participants feel is too long; 2. being bombarded by too many different surveys at once (over-surveying); and/or 3. filling out a survey that seems to ask irrelevant questions (survey disillusionment) (Porter, 2004; Porter, Whitcomb, & Weitzer, 2004; Sinickas, 2007; Adams & Umbach, 2012). Porter (2004) found that while survey length seems to be correlated with survey fatigue, the effect seems to be moderate. But Porter's study occurred before the age of the internet survey, which has increased over-surveying in general (Adams & Umbach, 2012). By increasing over-surveying, there seems to be less tolerance for long surveys and for irrelevant surveys, which has increased nonresponsiveness. While nonresponsiveness is a clear sign of survey fatigue, participant ambivalence is also problematic. Both of these fatigue issues can result in a skewed collection of data that no longer represents the population being studied.

The purpose of this study was to analyze different ways to shorten the survey by using a variety of techniques, including confirmatory factor analysis (CFA), principal component analysis (PCA), and cluster analysis. During the study, I sought to answer the following research questions:

1. Which method – Principal Component Analysis or Confirmatory Factor Analysis – presents the most compelling way to decide which questions to eliminate from the survey?

2. Was the cluster analysis results consistent with the results of the PCA and CFA or did the results propose different questions to eliminate from the survey?
3. Why and how would these methods differ in their resulting analyses?
4. Can the shortened survey then be used to perform linear regression in an attempt to build a model that might predict engineering student persistence?
    a. Which main effects are statistically significant in the analysis of the FACETS Design Identity & Beliefs survey data?
    b. Which factors contribute most significantly to the variation seen in the intent to persist average score and therefore a student's current willingness to continue within the engineering degree?

Methods

Study Design

Table 1 summarizes the plan for the statistical analysis overall. Each statistical model that requires an output needs the assumptions checked, a systematic way to find the most reduced model, a comparison between the reduced model and the full model, and a summary of the data both graphically and in tables. The statistical models that require no output still often require graphical analysis, checked assumptions, and results in tables.

Table 1. Summary of the statistical analysis performed for this project.

Indicators/Predictors (X)	Output/Dependent Variable (Y)
Descriptive Analysis of demographics listed in Table 5	No Output needed. Descriptive Analysis describes the data using measures of center and measures of spread.
CFA for latent variables listed in Tables 3 & 4	

PCA for latent variables listed in Tables 3 & 4	No Output needed. These techniques simply categorize predictors.
Cluster Analysis for latent variables listed in Tables 3 & 4	
Multiple regression and ANOVA for latent variables listed in Tables 3 & 4	IntPers (Intent to persist) AVERAGED
Multivariate regression and MANOVA for latent variables listed in Tables 3 & 4	IntPers (Intent to persist) AVERAGED

Participants and Setting

The study participants were University of New Mexico students enrolled in Chemical and Biological Engineering (CBE) courses from 2015-present. The specific dataset I used for the CFA, PCA, and Cluster Analyses included CBE students from Fall 2015, Fall 2016, Fall 2017, and Fall 2018. I exclusively included Falls in this analysis to consistently capture students starting their academic year. I included the original data (pre-tests and post-tests) from Fall 2015, Spring and Fall 2016, Spring and Fall 2017, and Fall 2018 for the regression analysis and ANOVAs because I wanted to build the model on the largest dataset I could clean. The study participants signed an informed consent form at the beginning of the semester of their first CBE class in the study (IRB 10915). Students who were analyzed were taking one of the four main Chemical and Biological Engineering (CBE) classes: Introduction to Chemical Engineering and Biological Engineering (CBE 101), Chemical and Biological Engineering Computing (CBE 253), Introduction to Transport Phenomena (CBE 311), and Chemical Engineering Design (CBE 493L).

I originally analyzed the data using descriptive analysis techniques. Descriptive analysis techniques allow us to measure the center and the spread of a dataset as well as the position of specific data points within the larger dataset. Measures of center (mean, mode, and median), measures of spread (range, inter-

quartile range, standard deviation, and variance), and measures of position (quartiles and z-scores) are not as relevant when analyzing categorical data and were therefore not included in Table 2. Instead we look at the frequencies of responses to determine which information is the most descriptive of the participants in this dataset. The results of the descriptive survey analysis are in Table 2.

Table 2. Descriptive statistics for the demographics survey data, including the variable, the possible responses and the values coded for the responses, the mode, the numbers of students who answered that question, the total responses, and the frequency %.

Variable	Values	Possible Responses	Mode	Number of Students	Total n responses	Frequency %
Home Language	1	Only/mostly English	1	217	329	**65.96%**
	0	Another language or languages AND English		57		17.33%
	-1	Only/mostly another language		55		16.72%
Gender	1	Male	1	186	328	**56.71%**
	0	Female		142		43.29%
Age	0	17 or younger	1	7	329	2.13%
	1	18-24		283		**86.02%**
	2	25-30		22		6.69%
	3	31-40		15		4.56%
	4	41 or older		2		0.61%
First Gen College	0	Yes	1	90	326	27.61%
	1	No		236		**72.39%**
Hispanic	0	Yes	1	141	329	42.86%
	1	No		188		**57.14%**
Race	0	American Indian or Alaska Native	1	14	325	4.31%
	1	White, not Hispanic/Latino		172		**52.92%**
	2	African America/Black		5		1.54%
	3	Asian		38		11.69%
	4	Native Hawaiian or Other Pacific Islander		5		1.54%
	5	Hispanic/Latino		68		20.92%
	6	Other/Mixed		5		1.54%
	7	Prefer not to state		17		5.23%

	8	Hispanic and American Indian		1		0.31%
Race_Contrast	0	Hispanic, Other non-white	1	82	204	40.20%
	1	White or Asian		122		**59.80%**
Urban_Call	0	Small town or surburban	0	181	329	**55.02%**
	1	Urban		148		44.98%
Mother_EdAtt	0	Less than High School	3	29	328	8.84%
	1	Completed High School		66		20.12%
	2	Some college or completed a 2-year degree		76		**23.17%**
	3	Completed a Bachelor's degree		78		**23.78%**
	4	Attended some graduate or professional school		9		2.74%
	5	Obtained a graduate or professional degree		70		21.34%
Father_EdAtt	0	Less than High School	3	30	319	9.40%
	1	Completed High School		57		17.87%
	2	Some college or completed a 2-year degree		60		18.81%
	3	Completed a Bachelor's degree		83		26.02%
	4	Attended some graduate or professional school		14		4.39%
	5	Obtained a graduate or professional degree		75		23.51%
Econ_Stat	0	Low	2	33	327	10.09%
	1	Lower middle		54		16.51%
	2	Middle		145		**44.34%**
	3	Upper middle		91		27.83%
	4	High		4		1.22%
Eng_Any	0	No relative	1	107	329	32.52%
	1	Any relative		222		**67.48%**
HS_Calc	0	Did not take	1	115	234	49.15%
	1	Did take		119		**50.85%**
HS_Chem	0	Did not take	1	63	245	25.71%
	1	Did take		182		**74.29%**
HS_Phys	0	Did not take	1	101	237	42.62%

	1	Did take		136		**57.38%**

We can see from the descriptive statistics, particularly the modes and the frequencies, shown in Table 2, several important points. The home language of most students is predominantly English, although 30+% percentage of students predominantly speak another language at home. The gender of the study participants skews slightly male. Most students fall in the age range of 18-24 years old. Most students are not first-generation college students, but a small, important population is first-generation. Students in the CBE program skew white, but 20+% of the population is Hispanic students. This finding makes sense as the University of New Mexico (UNM) is also a Hispanic-serving Institution. There is also a fairly large population of students from other underserved groups. Many students are from small towns or suburban areas. Most students in the study have parents who both went to college and possibly received a bachelor's degree. Most students in the study are middle class and more than two-thirds of them have a relative who is an engineer. More than half of the students in the study have had calculus and physics in high school, and nearly three-quarters of the same students took chemistry in high school.

Data Collection

The FACETS grant proposal (NSF Award number 1623105) discussed implementing the full survey, including student demographics and information on student experience with design and self-efficacy beliefs, at the beginning of Chemical and Biological Engineering (CBE) 101 and 251. A shortened survey,

which did not include demographics, would be implemented as a posttest at the end of each course. Baseline data was collected in Fall 2015, and this statistics project includes that data. The variables for the survey are defined by their corresponding survey question in Table 4.

Table 3 shows the variables in the dataset, including an expanded variable name, an explanation of the variable name, and general categories (Design beliefs and knowledge) in which the variables belong. The three starred rows show variables that were added to the survey by faculty amid the study. While Prep was left out of this statistical analysis study due to a lack of consistency in the question and its use in the surveys and Team was only used in the Regression Analysis, DesChalMot was included all of the statistical analyses.

Table 3. Variables Legend. This table acts as a legend for the variables, what they describe in short form, and a longer general description of each variable.

Variable Name	Expanded Variable Name	Explanation of Variable Name
Design beliefs and knowledge		
MeetNeeds	Meets Needs	Does the design meet the needs of the client?
IllStruc	Ill Structured	Design is an ill structured activity.
Iter	Iteration	Design is iterative.
Creativ	Creativity	Design is creative.
Frame	Design Frame	Framing design is an important aspect of a design process.
Other factors salient to design outcomes		
DesSelfEff	Design Self Efficacy	These questions probe students' self-efficacy for designing.
IntPers	Intent to Persist	These questions probe students' intent to persist in an engineering career.
Social	Social Integration	These questions probe the social aspects of students' engineering major and career.
DegChoi	Degree Choice	Is student's degree choice engineering?
ProfIden	Professional Identity	These questions probe students' ability to identify with engineering professionals.

Prep*	Preparation for Engineering Coursework	This question asks about students' prior preparation for engineering coursework.
Team*	Willingness to be on a Team	These questions probe teamwork in engineering coursework and careers.
DesChalMot*	Design Challenge Motivation	These questions probe students' motivation for design challenges.

***Items added by faculty involved in the study**

The gray shaded rows in Table 3 distinguish those variables that had reverse Likert scale answers and are also designated with an "_R" at the end of the variable (after the year and month are stated). Based on student feedback, the question and variable shown in red was determined to be a confusing question and was therefore eliminated from the analysis. The blue shaded rows show variables and questions that were added to the survey starting in 2017.

Table 4. The variable names and survey questions for the Likert scale section of the FACETS Design Identity & Beliefs survey. This table was referenced when data cleaning and during the analysis. All items used a five-point Likert scale from strongly agree (1) to strongly disagree (5) as question responses. Note that the variables with an "_R" at the end of the variable name have a reverse Likert scale from the other questions on the survey.

Variable name	Question text
MeetNeeds1	In design, a primary consideration throughout the process is addressing the question "Who will be using the product?"
MeetNeeds2	Design is the process of devising a system, component or process to meet a desired need.
MeetNeeds3	Design begins with the identification of a need and ends with a product or system in the hands of a user.
IllStruc1	In design, the problem and the solution co-evolve, where an advance in the solution leads to a new understanding of the problem.
IllStruc2 _R	Design problems have right answers
IllStruc3	Design problems have multiple possible solutions and multiple ways to get to the solution
IllStruc4 _R	Designers of equal skill and experience should come to the same design solution given the same initial design problem
IllStruc5 _R	An expert designer is usually right on the first try when designing
Iter1	Design is iteration
Iter2 _R	Design is usually a linear, predictable process
Iter3 _R	Design is a goal-oriented, constrained activity
Creativ1 _R	Expert designers typically consider many possible ideas which leads to better solutions
Creativ2 _R	Constraints typically hinder creative design
Creativ3	Creativity is integral to design. Every design project involves creativity.

Frame1	Design is as much a matter of finding problems as it is of solving them.
Frame2 _R	The design problem is framed by the client or customer, then solved by the designer
Frame3	Design, in itself, is a learning activity where designers continuously refine and expand their knowledge.
DesSelfEff1	I am confident I could develop possible design solutions to an authentic engineering design problem
DesSelfEff2	I am confident I could select the best possible design for an authentic engineering design problem
DesSelfEff3	I am confident I could construct a prototype for an authentic engineering design problem
DesSelfEff4	I am confident I could evaluate and test a design solution to an authentic engineering design problem
DesSelfEff5	I am confident I could describe the work professional engineers do.
DesSelfEff6	I am confident I could identify a need in an authentic engineering design problem
IntPers1	I intend to complete a major in Chemical engineering
IntPers2	I intend to complete a major in engineering other than Chemical engineering
IntPers3 _R	I have considered pursuing a major outside of engineering in the past few months.
IntPers4	After graduation, I plan to go to graduate school in an engineering discipline
IntPers5	I plan to pursue a career in engineering
Social1	I belong to a professional engineering organization, such as the Hispanic Engineering and Science Organization, the American Indian Science and Engineering Society, the National Society of Black Engineers, AIChE, BMES, or the Society of Women Engineers.
Social2	I participate in engineering-related activities outside coursework
Social3 _R	Most of my friends and social interactions are outside of engineering
Social4	The faculty and staff make engineering feel like a welcoming place for me
Social5 _R	It is very important to me to be involved in non-engineering activities, such as hobbies, civic or church organizations, campus publications, student government, social fraternity or sorority, sports, etc.
DegChoi1 _R	My family or friends have encouraged me to pursue a degree outside of engineering
DegChoi2	My primary reason for pursuing engineering as a career is because a parent, guardian, teacher or guidance counselor encouraged me to do so.
DegChoi3	My parents want me to be an engineer
DegChoi4	My parents would disapprove if I chose a major other than engineering
DegChoi5	Before college, I had a lot of knowledge about the engineering profession
DegChoi6	My prior academic experiences have prepared me to be successful in engineering
ProfIden1	I feel like an engineer
ProfIden2	I am familiar with what a practicing engineer does.
ProfIden3	The main reason I considered engineering as a major is that I know what engineers do and the work appeals to me
ProfIden4	I participated in some type of engineering internship, club, course, or camp prior to university
ProfIden5	I am confident that I can succeed as an engineering major
ProfIden6	Creative thinking is one of my strengths
ProfIden7	I am skilled at solving problems that have multiple possible solutions.
Prep	My past experiences are relevant in my engineering coursework.
Team1	Teamwork is important in engineering.
Team2 _R	I prefer to work by myself rather than in a team.
DesChalMot1	I would be motivated to work on a design challenge if I thought the design could help people.
DesChalMot2	I would be motivated to work on a design challenge if I thought the design could help the environment or result in a more sustainable/green solution.

DesChalMot3	I would be motivated to work on a design challenge if I thought the design could be highly innovative and novel.

Table 5. The demographics variable names, survey questions, and question responses for the FACETS Design Identity & Beliefs survey.

Demographics		
HS_Calc	Which of the following did you complete in high school?	Did not take (0) vs. Did take (1)
HS_Chem	Which of the following did you complete in high school?	Did not take (0) vs. Did take (1)
HS_Phys	Which of the following did you complete in high school?	Did not take (0) vs. Did take (1)
Home Language	Growing up, what language or languages were spoken in your home	Only/mostly another language (-1); Another language or languages AND English (0); Only/mostly English (1)
Gender	Gender	Female (0); Male (1)
Age	Age	17 or younger (0); 18-24 (1); 25-30 (2); 31-40 (3); 41 or older (4)
First_Gen_Coll	Are you a first generation college student?	Yes (0) or No (1)
Hispanic	Are you Hispanic or Latino?	Yes (0) or No (1)
Race	What is your race? Select one or more:	American Indian or Alaska Native (0); White, not Hispanic/Latino (1); African American/Black (2); Asian (3); Native Hawaiian or Other Pacific Islander (4); Hispanic/Latino (5); Other/Mixed (6); Prefer not to state (7); Hispanic and American Indian (8)
Race_Contrast	What is your race? Select one or more:	Hispanic, Other non-white (0); White or Asian (1)
Urban_Call	Which best describes where you lived before attending college?	Small town or surburban (0); Urban (1)
Mother_EdAtt	Please indicate the highest level of education attained by your mother	Less than high school (0); Completed high school (1); Some college or completed a 2-year (e.g., associates) degree (2); Completed a bachelor's degree (3); Attended some graduate or professional school (4); Obtained a graduate or professional degree (5)
Father_EdAtt	Please indicate the highest level of education attained by your father	Less than high school (0); Completed high school (1); Some college or completed a 2-year (e.g., associates) degree (2); Completed a bachelor's degree (3); Attended some graduate or professional school (4); Obtained a graduate or professional degree (5)
Econ_Stat	Would you describe your family as low, lower middle, middle, upper-middle, or high income?	Low (0); Lower middle (1); Middle (2); Upper middle (3); High (4)
Eng_Any	Do you have any family or close friends who are/were engineers? Check all that apply:	No relative (0); Any relative (1)
Acad_Stand	What is your current academic standing?	Freshman (1); Sophomore (2); Junior (3); Senior (4); 5th year senior (5); Grad student (6); Other (7)

Hours_Work	How many hours do you work in a typical week during the school year to earn money for yourself and/or your family	0 Hours (0); up to 5 (1); 6-10 (2); 11-15 (3); 16-20 (4); 21-25 (5); 26-30 (6); 31-35 (7); 36-20 (8); More than 40 (9)
GPA	What is your current cumulative grade point average?	As percent (continuous)
Major	What is your (intended) major?	Other science, math or technology field (-1); Other engineering (0); Chemical engineering (1)
Intern	Have you participated in any type of engineering internship in the past year? If so, please briefly describe it. If not, leave blank.	None (0) or Any (1)
EngHS	Have you ever participated an engineering activity prior to college?	None (0) or Any (1)

Data Cleaning

The original Design Identity & Beliefs survey dataset, which included Fall and Spring of 2015 and 2016, was already cleaned. I undertook cleaning the survey data to increase the dataset we could use. Data cleaning is a process that makes the data valid for statistical analysis and may involve reassigning values, relabeling text as numbers, or reorganizing the data. The data cleaning required several steps, including:

1. Downloading the csv (comma separated values) data from a google form for each semester the survey was administered.
2. Keeping the csv file with the survey dataset on a locked and password-protected computer and deleting the identifiable data once the data has been cleaned (and de-identified).
3. To data clean in Excel, the question responses for each variable (column) had to be re-coded using the numerical values provided in Table 5. This process took finding and replacing the written response (such as Male for Gender) with a numerical value (such as 1). This step shows the very human

side of taking a survey; some participants marked multiple answers, some wrote in answers that weren't coded, etc. I felt it was important to get to know this data thoroughly and to have double-checked the input, row by row.

4. Once the data had been re-coded, the students had to be double-checked against a master list of students who had consented to participate in the study and their study IDs. Study IDs for the consented students were then added to the Excel spreadsheet and the names of the survey respondents were deleted. All students who had not consented to have their data used in the study were deleted as well.
5. Reorganizing the data into different forms (longitudinal tracking, demographics, latent variables including DesSelfEff, IntPers, Social, and ProfIden) so that the different statistical analysis could proceed in R efficiently.
6. To eliminate blanks and to balance the dataset, a dummy value (-2) was inserted into the excel spreadsheet and then eliminated in the R code at the beginning of the analysis.

The dataset included the original data (pre-tests and post-tests) from Fall 2015, Spring and Fall 2016, Spring and Fall 2017, and Fall 2018. The number of student participants had doubled.

Data Analysis

Graphing the Data

The first step in any model analysis is to plot the data. Scatterplots for such a large set of variables are a bit hard to view but are important to incorporate within the analysis. The goal of plotting the data is to look for patterns in the data.

The first plot (Figure 1) has random scatterplots with no color variation on the right side of the diagonal and correlation values between variables on the left side of the diagonal. The correlation values are above 0.70 for DesSelfEff1 and DesSelfEff3 (0.71), DesSelfEff3 and DesSelfEff4 (0.77), DesSelfEff1 and DesSelfEff6 (0.71), ProfIden6 and ProfIden7 (0.66), and the averages highly correlate with the questions they've averaged. The highest correlation is between DesSelfEff3 and DesSelfEff4, which makes sense because developing a prototype (a potential design solution), testing, and evaluating a design solution seem to be linked.

The second plot (Figure 2) only includes the variables used in the Regression Analysis and ANOVA (Semester, DesSelfEffAVG, IntPersAVG, SocialAVG, ProfIdenAVG). It has the density functions on the right side of the diagonal and the random scatterplots (color based on the semester) on the left side of the diagonal. The diagonals in each case are line charts for the data to show the overall pattern for each variable.

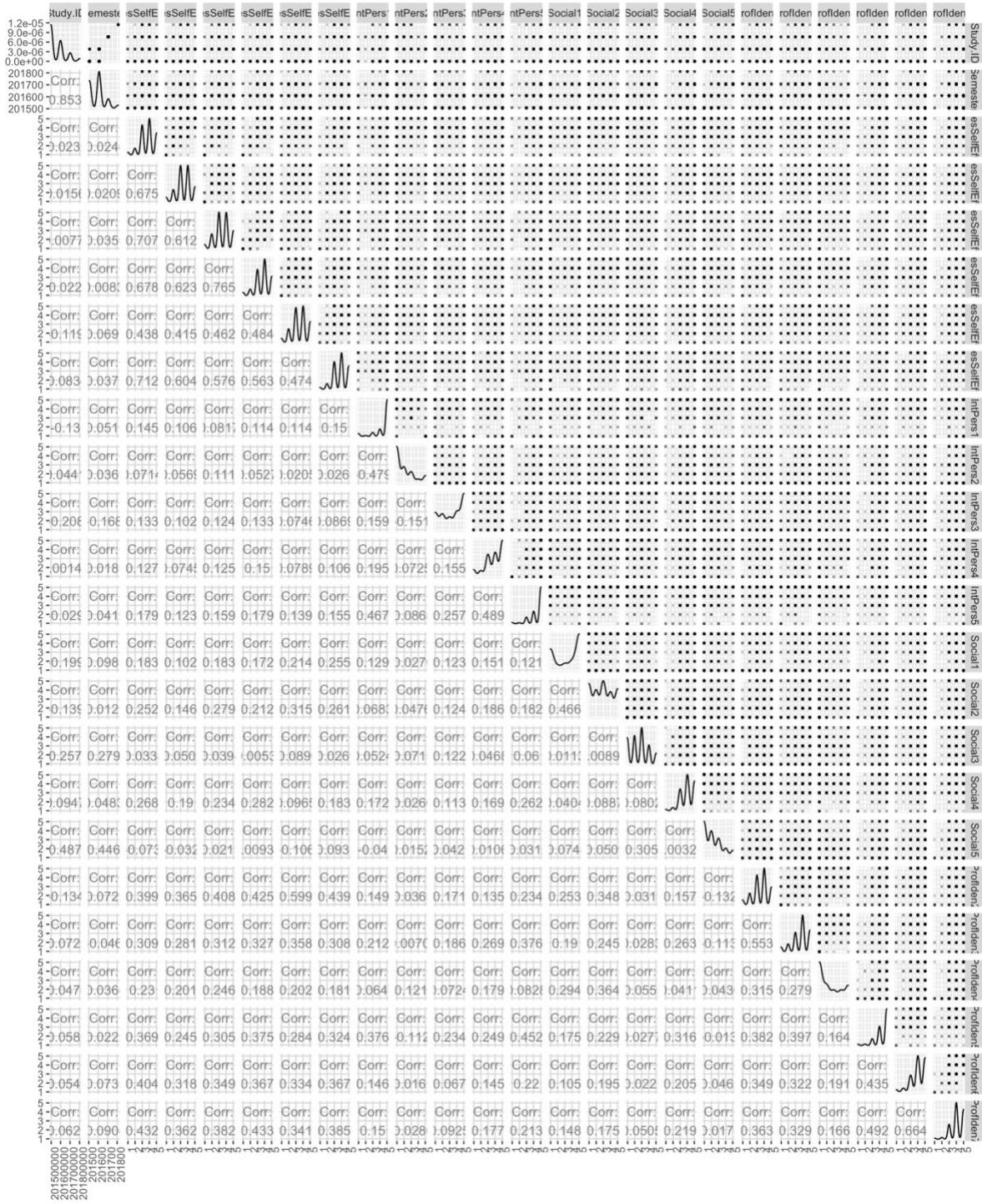

Figure 1. This plot has random scatterplots with no color variation on the right side of the diagonal and correlation values between variables on the left side of the diagonal. The diagonal is the smoothed histogram for each variable to show the overall pattern.

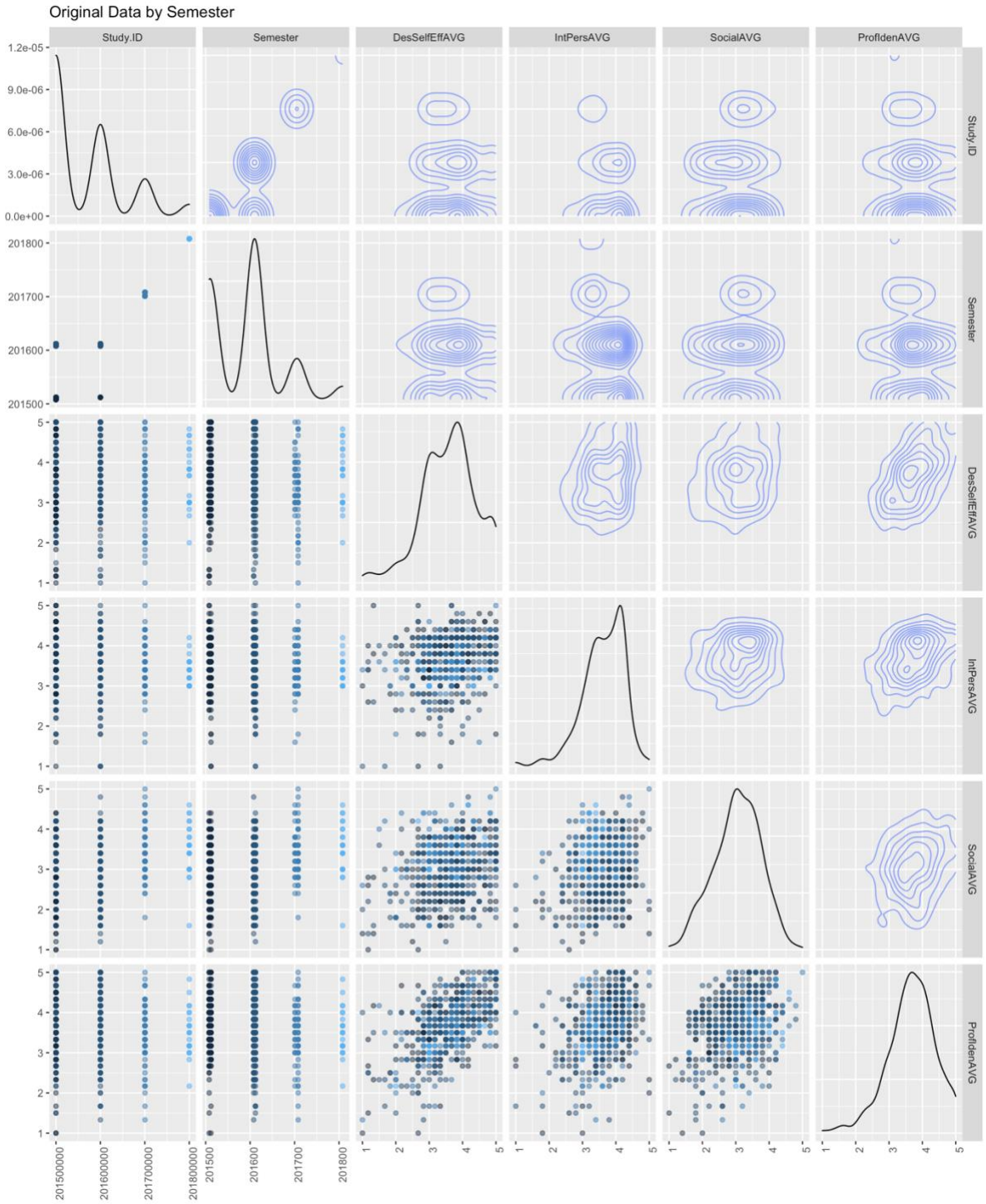

Figure 2. This plot has the density functions are the right side of the diagonal and the random scatterplots (color based on semester) on the left side of the diagonal for the variables included in the Regression Analysis only (Semester, DesSelfEffAVG, IntPersAVG, SocialAVG, ProfIdenAVG). The

diagonal is the smoothed histogram for each variable to show the overall pattern.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) was proposed to confirm a certain factor structure regarding The Survey questions. In general, factor analysis can be used to condense variables and/or expose relationships between clusters of responses. The reason why factor analysis is used is "to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called [latent] *factors*" (Johnson & Wichern, 2013, p. 481, parenthesis mine). Factor analysis was used in this project to try to eliminate extraneous questions in the survey to shorten the survey length.

Each CFA analysis used a Model Chi-Square, Comparative Fit Index (CFI), a Root-Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). A chi-square test between the orthogonal model and the proposed model to test the relative fit of the model was also run on each factor analysis. The Model Chi-Square test value assesses the overall fit of the proposed model against the null, which states that the model fits perfectly. The Model Chi-Square test is sensitive to sample size and shows the discrepancy between the sample and the fitted covariance matrices. The Comparative Fit Index (CFI) compares the user (proposed) model with a stricter baseline (null) model, which usually means that all of the variables in the model have variation but no correlation, through the formula (Kenny, 2015):

$$CFI = \frac{(\chi^2 - df)_{Null\ Model} - (\chi^2 - df)_{Proposed\ Model}}{(\chi^2 - df)_{Null\ Model}}$$

Generally, if the CFI is higher, it's considered a better model fit. Any model with a CFI of greater than 0.9 is considered an ok fit. The Tucker-Lewis Index (TLI) is similar to the CFI but is more conservative because it penalizes overly complex models. Both CFI and TLI rely on average correlations of data, and if the average correlations amongst the variables are not high, then these measures will not be high. This study will only report the CFI.

The Root Mean Square Error of Approximation (RMSEA) is an absolute measure of fit, which means that the best fitting model has a fit of zero. Typically, RMSEA is calculated as (Kenny, 2015):

$$RMSEA = \frac{\sqrt{(\chi^2 - df)}}{\sqrt{df(N-1)}}$$

where N is the sample size and df are the degrees of freedom in the model. The RMSEA value, therefore, shows how far the proposed model is from the best model and the smaller the RMSEA value, the better the fit. A 90% confidence interval can be calculated for the RMSEA and should typically range between 0.05 and less than 0.08 if the model fit is good. (Kenny, 2015)

The Standardized Root Mean Square Residual (SRMR) is the standardized difference between the observed correlation and the fitted correlation. The SRMR is an absolute measure of fit, like RMSEA, which means that the best fitting model has a fit of zero. Any value less than 0.08 is considered a good fit and it does not penalize the complexity of the model. (Kenny, 2015)

The chi-square test between the orthogonal model, which is the model that does not allow covariances between the latent factors, and the proposed model, which is the model that does allow covariances between latent factors, is used to

test the relative fit of the model as well. The orthogonal model is a simpler model because the latent factors are considered independent, and the chi-square difference test between the two models is administered through the ANOVA function. A p-value less than a standard alpha value (0.05) on this test rejects the null hypothesis, which states that the simpler model is the better fit, and therefore latent variables can be treated independently.

The confirmatory factor analysis was performed on student data collected during the Fall semesters of 2015-2018. Only fall semesters were chosen to normalize the data; students from Fall vs. Spring semesters can be rather different.

Principal Component Analysis

Both factor analysis and principal components analysis feature extraction techniques that are used to explain large sets of correlated multivariate data by mainly combining variables into latent variables, which thereby reduces the total number of variables used in the analysis. The differences between factor analysis (FA) and principal components analysis (PCA) lie in how the variables are combined. PCA recombines the variables using linear combinations of the original independent variables to form new variables (sometimes called latent variables). These new variables are created by multiplying Z [the centered (possibly standardized) version of X] by the eigenvectors from the covariance matrix (Z^TZ), which means that the new variables contain information on how the original variables were associated with one another, the directions in which the data was scattered, and the relative importance of the directions (which allows us to rank the

new PCA variables) (Brems, 2017). The new PCA variables are orthogonal and independent to one another yet are also less interpretable. FA uses regression analysis and "loads" the factors with pre-determined betas, or the correlation of the item with the factor (i.e. $Y_n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$). Table 6 describes the major differences between factor analysis and principal component analysis and is paraphrased from Everitt & Dunn, 2001 (p. 287):

Table 6. A summary of the major differences between factor analysis vs. principal component analysis. While factor analysis is based on linear regression, principal component analysis has no overall model for the data. Paraphrased from Everitt & Dunn (2001, p. 287).

Factor Analysis	Principal Components Analysis (PCA)
Hypothesizes a model for the data	No model for the data involved
Tries to explain covariances or correlations of observed variables through a few common factors	Explains the variance of observed variables
If the number of factors (m) changes, even by 1 (m+1), it can affect the entire analysis.	If the number of factors (m) changes, the first m principal components remain unchanged.
For maximum likelihood factor analysis, the results of analyzing the correlation matrix or the covariance matrix or the factor analysis itself are essentially equivalent.	No relationship exists between the principal components and the correlation matrix or the covariance matrix for the sample.

The factor analysis was completed using confirmatory factor analysis (CFA), which differs from exploratory factor analysis (EFA) in that the number of factors is already known and the loadings on those factors are preset. We would expect that items that relate to the same factor (latent variable) would be highly correlated. We would also expect that items for different factors would not be correlated.

I used Principal Component Analysis (PCA) as a secondary method to confirm (and possibly expand) the results of the CFA. Dr. Svihla requested the PCA

because the PCA would show variation in a way that batched the questions more thoroughly.

Principal Component Analysis is a method used to describe the variation in a set of multivariate data by building linear combinations of the observed variables to make components. The components are derived in decreasing order of importance in terms of explaining the variance within the dataset. In other words, the first principal component explains as much variation in the original dataset as possible; the second principal component accounts for as much of the remaining variance that is now uncorrelated with the first component as possible and so on.

Cluster Analysis

Cluster analysis was used to approach the goal of shortening the number of survey questions in a new way. What if we did not know the survey content, but knew that our client wanted a shorter survey that resembled the original survey as closely as possible in terms of content and the data collected? Having no prior knowledge of the survey content means that we would need to approach the shortening of the survey in a whole new way. Because there is no analysis of group classifications before cluster analysis is used, cluster analysis asks major questions like how groups can be formed from inter-subject similarities and weighted measures, and then once the groups are formed, what kinds of weighted measures are reported and what can we infer from their relative statistical significance?

Cluster analysis is used to categorize questions or responses (i.e. objects) in such a way as to maximize the inter-group distances and minimize the intra-group distances. Distance is often measured as a Euclidean distance,

$$d = \sqrt{\sum_{i=1}^{N} (cluster_1 - cluster_2)^2}$$

and is used to show whether individual objects in the cluster analysis are similar or dissimilar. Most of the cluster analysis I used was model-based, which means each cluster had its own model and the point of the analysis was to find the best fit model for all of the clusters. Most of this model-based cluster analysis was shown through density plots, which means it used density functions to measure the connectivity and the similarity between objects.

Clusters were analyzed using k-means analysis and a normal finite mixture model fitted by an expectation-maximization (EM) algorithm. Both methods of cluster analysis were performed on the same dataset as used in the confirmatory factor analysis and the principal component analysis. Kmeans cluster analysis uses an unsupervised machine learning algorithm to analyze a dataset repeatedly while the algorithm sorts the data into a specified number of clusters starting from random assignments. Unsupervised simply means the outcome is not predetermined or predicted in advance. Clusters in k-means have a spherical shape; each iteration involves sorting the observation into a cluster and recalculating the centroid mean of the cluster. Once the within-cluster variation (calculated as the sum of the Euclidean distance between the data point observations and their centroid mean) cannot be reduced further, the algorithm ceases (Kodali, 2015).

Overall, k-means analysis is more simplistic and requires an input of the number of groups desired. I decided the number of groups would be twelve, based on the number of desired latent variables shown in Table 3 with the Prep variable removed. The distances between clusters in k means are calculated via a sum of squares. When k-means was run on the pre-test data, the ratio of the between-cluster sum of squares to the total sum of squares was 29.1%, which is low. The ratio accounts for the amount of total sum of squares of the data points between the clusters. When kmeans was run on the post-test data, the ratio of between-cluster sum of squares to total sum of squares was 31.9%, which is still low. While we would want to increase these values, we also don't want to overfit the data.

We decided that clustering the participant responses to the survey questions by participant (study ID) might help us decide which questions to eliminate by highlighting the lone wolf questions in particular. Repeating the clustering algorithm using different permutations would allow us to view the lone wolf questions from multiple angles. For the cluster analysis, I chose to use k-means cluster analysis and cluster analysis performed through *mclust*.

The *mclust* R package employs finite normal mixture modeling that is fitted by an expectation-maximization algorithm for maximum likelihood estimation. The R package *mclust* intuitively performs model-based clustering analysis and dimension reduction by applying maximum likelihood estimation and Bayes criteria to identify both the most likely model and the most ideal number of clusters. It uses hierarchical clustering for normal mixture models to find the most optimal model via EM (Expectation-Maximization).

Normal finite mixture modeling assumes that there are *n* independently identical distributed observations and *x* is a sample of *n.* Every individual observation (*x*) has its own distribution, which is a probability density function derived from a finite mixture model of *G* mixture components, given by (Scrucca, Fop, Murphy, & Raftery, 2016):

$$f(\boldsymbol{x_i}; \boldsymbol{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(\boldsymbol{x_i}; \boldsymbol{\theta_k})$$

Where $\boldsymbol{\Psi} = \{\pi_1, \ldots, \pi_{G-1}, \theta_1, \ldots, \theta_G\}$ are the parameters of the mixed model and $f_k(\boldsymbol{x_i}; \boldsymbol{\theta_k})$ is the component density for x_i with parameter vector θ_k with a mixed weight or probability loading factor (π_k). (Scrucca et. al, 2016)

The output is given as ten parameterized covariance structures with density estimation and other graphical representations. Table 7 shows the parameterizations of the within-group covariance matrix in the *mclust* package as well as the corresponding geometric characteristics, which are also determined by the covariance matrix.

Table 7. Parameterizations of the covariance matrix available for hierarchical clustering (HC) or EM for multidimensional data (Fraley et. al, 2012, p. 8). The model column involves scalars (λ_k, which controls the volume of the ellipsoid), the identity matrix (*I)*, diagonal matrices which specify the shape of the density contours ($\boldsymbol{A}_k$), and orthogonal matrices which determine the orientation of the ellipsoid ($\boldsymbol{D}_k$) (Scrucca et. al, 2016).

Identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
E		•	•	(univariate)	equal		
V		•	•	(univariate)	variable		
EII	λI	•	•	Spherical	equal	equal	NA
VII	$\lambda_\kappa I$	•	•	Spherical	variable	equal	NA
EEI	λA		•	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_\kappa A$		•	Diagonal	variable	equal	coordinate axes

EVI	λA_κ		•	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_\kappa A_\kappa$		•	Diagonal	variable	variable	coordinate axes
EEE	λDAD_T	•	•	Ellipsoidal	equal	equal	equal
EEV	$\lambda D_k AD_{kT}$		•	Ellipsoidal	equal	equal	variable
VEV	$\lambda_\kappa D_k AD_{kT}$		•	Ellipsoidal	variable	equal	variable
VVV	$\lambda_\kappa D_k A_k D_{kT}$	•	•	Ellipsoidal	variable	variable	variable

The normal finite mixture model is fitted by an expectation-maximization (EM) algorithm for maximum likelihood estimation.

Regression Analysis

The main research questions are formulated in a regression model and answered based on model fit, the significance of specific factors to the model, and whether the assumptions for that model are violated.

The regression analysis outcomes included the following outcome variables: intent to persist, professional identity, and scores on the design challenges for innovation. I analyzed the many variables on the survey against Intent to Persist, a summed output of individual semester and student Likert scale scores. This statistical analysis was performed using multiple regression (individually) and multivariate regression techniques.

The full model generalized formula for latent analysis is:

$$Y_{ijkmn} = \mu_{...} + \alpha_i + \beta_j + \gamma_k + \delta_m + \varepsilon_{ijkmn}$$

Where i, j, k, m are the indices for the number of latent variables analyzed. The output variable (Y) is the averaged latent variable Intent to Persist (or IntPers). The latent variables analyzed as predictors included DesSelfEff, Social, ProfIden, and, even though it was not a latent variable, Semester was included as well as a

factor variable. These variables correspond to α, β, γ, and δ respectively, and n is the number of observations for all variables analyzed. The overall mean for all factors is $\mu_{...}$ and ε_{ijkmn} symbolizes the error not accounted for by the model ($\varepsilon_{ijkmn} = Y_{ijkmn} - \bar{Y}_{ijkm}$).

The assumptions for this model ($\varepsilon_{ijkmn} \overset{iid}{\sim} N(0, \sigma^2)$) include the following:

1. The variance (σ_2) is constant for all treatments as well as the error.
2. The observations are collected independently.
3. The error term residuals are normal and identically distributed.

The full additive multiple regression model, using the averaged columns for IntPers (Intent to Persist), DesSelfEff (Design Self Efficacy), Social, ProfIden (Professional Identity), and Semester as a factor, was created and is shown below:

$$IntPers_{ijkmn} = \beta_0 + \beta_1 DesSelfEffAVG + \beta_2 SocialAVG + \beta_3 ProfIdenAVG + \beta_4 Semester + \varepsilon_{ijkmn}$$

The analysis was performed using R (ver. 3.6.1 for Mac OS X) and RStudio (ver. 1.2.5019) with the following packages: gglot2, GGally, lavaan, knitr, dplyr, tidyr, devtools, ggbiplot, FactoMineR, factoextra, and mclust.

Results

Confirmatory Factor Analysis (CFA)

As I've stated earlier in this paper, Dr. Svihla needed to shorten the original survey as she believed respondents were getting survey fatigue from answering such a lengthy survey. The research questions for this part of the study were:

1. Which questions can we remove because they don't cluster with others as we expect?

2. Which questions don't cluster with the other questions asking about the same latent variable?
3. Which questions are highly variable? Which questions aren't variable in that they don't change at all from pre- to post-test or from semester to semester?

CFA Individual Question Analyses by Pre-Test vs. Post-Test

This section will discuss each latent variable or theme in the survey with respect to clustering and intercorrelatation between items. The tables provided in each section have several values and are defined as: B's are the factor loadings, which can be interpreted like a regression coefficient, but are not the same thing (this is why the regression coefficient Beta is also stated in the table); SE is the Standard Error for each estimated parameter; Z is the Wald statistic (B divided by SE), assuming this CFA meets the assumption of normality; p-value is the p-value, which tests against the null hypothesis that the parameter equals zero in the population; Beta is the std.all, or a standardized regression coefficient, which is stated as β (the parameter value) within the linear regression CFA model. The significance stars simply show a significance comparison between the p-value listed and alpha values of 0.05, which yields one star if the p-value is lower than 0.05; 0.01, which yields two stars if the p-value is lower than 0.01; and 0.001, which yields three stars if the p-value is lower than 0.001. If the question or item is significant, then the question/item is important to describing the latent variable, and it clusters and intercorrelates with the other significant questions.

Meets Needs *(MeetNeeds)* - Does the design meet the needs of the client?

As shown previously in Table 4, the MeetsNeeds questions probe the degree to which students view design as involving meeting client or customer needs. In analyses of both the pre-tests and the post-tests shown in Table 8, the MeetsNeeds questions all show significance, which means that all of the MeetsNeeds questions are clustered, important in terms of measuring the latent variable, and intercorrelate with one another.

Table 8. MeetsNeeds CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
MeetNeeds	MeetNeeds1	0.202	0.047	4.274	0.000	0.312	***
MeetNeeds	MeetNeeds2	0.517	0.075	6.886	0.000	0.777	***
MeetNeeds	MeetNeeds3	0.561	0.085	6.621	0.000	0.690	***
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
MeetNeeds	MeetNeeds1	0.322	0.068	4.748	0.000	0.452	***
MeetNeeds	MeetNeeds2	0.381	0.063	6.029	0.000	0.638	***
MeetNeeds	MeetNeeds3	0.664	0.095	6.954	0.000	0.827	***

Ill-Structured *(IllStruc)* - Design is an ill-structured activity.

As shown previously in Table 4, the IllStruc questions probe the degree to which students see design problems as ill-structured. Table 9 shows that while all of the IllStruc questions were significant in the pre-tests, some showed less significance in the post-tests, including questions 1, 3, and 4. Since the questions were significant overall (but differed slightly in the amount of significance), probably

all of the IllStruc questions cluster, are important in terms of measuring the latent variable, and intercorrelate with one another.

Table 9. IllStruc CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
IllStruc	IllStruc1	0.169	0.051	3.329	0.001	0.265	***
IllStruc	IllStruc2_R	-0.482	0.087	-5.570	0.000	-0.434	***
IllStruc	IllStruc3	0.217	0.042	5.135	0.000	0.403	***
IllStruc	IllStruc4_R	-0.719	0.088	-8.152	0.000	-0.680	***
IllStruc	IllStruc5_R	-0.524	0.076	-6.865	0.000	-0.551	***
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
IllStruc	IllStruc1	0.234	0.104	2.255	0.024	0.338	*
IllStruc	IllStruc2_R	-0.498	0.132	-3.784	0.000	-0.441	***
IllStruc	IllStruc3	0.303	0.099	3.047	0.002	0.484	**
IllStruc	IllStruc4_R	-0.359	0.114	-3.140	0.002	-0.324	**
IllStruc	IllStruc5_R	-0.594	0.138	-4.315	0.000	-0.696	***

Iterative *(Iter)* – Design is iterative and Creative *(Creativ)* – Design is creative.

Table 4 has previously shown that the Iter questions probe the degree to which students see iteration as required within design. The Creativ questions probe the degree to which students understand the centrality of creativity within design. None of the Iter or Creativ questions seem to cluster together nor do they seem to contribute to the latent variables in either the pre-tests or post-tests in Table 10. This result seemed a bit weird - two entire latent variables (six total questions) were not significant?

Table 10. Iter and Creativ CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	Beta	Latent Factor	Indicator	B	Beta
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
Iter	Iter1	-0.010	-0.009	Creativ	Creativ1_R	See Table 10	
Iter	Iter2_R	-0.015	-0.017	Creativ	Creativ2_R		
Iter	Iter3_R	-20.128	-18.556	Creativ	Creativ3		
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
Iter	Iter1	0.004	0.004	Creativ	Creativ1_R	0.875	1.337
Iter	Iter2_R	0.005	0.006	Creativ	Creativ2_R	-0.070	-0.060
Iter	Iter3_R	48.698	43.749	Creativ	Creativ3	0.269	0.387

Looking at the Creative pre-test questions CFA analysis in Table 11, each of the questions did not contribute to the overall latent variable (creativity in design) nor did they cluster together as they should.

Table 11. Creativ CFA Individual Question Analyses by Pre-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Creativ	Creativ1_R	0.715	0.542	1.319	0.187	1.115	
Creativ	Creativ2_R	-0.092	0.097	-0.948	0.343	-0.085	
Creativ	Creativ3	0.282	0.217	1.299	0.194	0.427	

Possibly the weird result in Iter was due to a violation of normality in the individual tests for both the pre- and post-test data. However, the assumption of normality was not violated in either of the Creativ individual tests for the pre-or post-test data, the indicators (or questions) were simply not significant. Table 11 shows the pre-test data for Creativ.

To test for significance in a different way, I grouped the analysis of Creativ and Iter with other variables that I already knew to be significant (DesSelfEff and IllStruc) in Table 12.

Table 12. Grouped Analysis of DesSelfEff and IllStruc with Creativ and Iter

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
IllStruc	IllStruc1	0.324	0.083	3.910	0.000	0.467	***
IllStruc	IllStruc2_R	-0.362	0.132	-2.739	0.006	-0.320	**
IllStruc	IllStruc3	0.365	0.068	5.405	0.000	0.582	***
IllStruc	IllStruc4_R	-0.341	0.111	-3.086	0.002	-0.308	**
IllStruc	IllStruc5_R	-0.509	0.096	-5.295	0.000	-0.597	***
DesSelfEff	DesSelfEff1	0.717	0.056	12.700	0.000	0.864	***
DesSelfEff	DesSelfEff2	0.747	0.067	11.158	0.000	0.791	***
DesSelfEff	DesSelfEff3	0.801	0.075	10.679	0.000	0.771	***
DesSelfEff	DesSelfEff4	0.788	0.059	13.252	0.000	0.886	***
DesSelfEff	DesSelfEff5	0.479	0.068	7.007	0.000	0.555	***
DesSelfEff	DesSelfEff6	0.589	0.055	10.766	0.000	0.776	***
Creativ	Creativ1_R	0.423	0.060	7.060	0.000	0.680	***
Creativ	Creativ2_R	-0.116	0.114	-1.023	0.306	-0.098	
Creativ	Creativ3	0.491	0.068	7.244	0.000	0.709	***
Iter	Iter1	-0.066	0.061	-1.085	0.278	-0.074	
Iter	Iter2_R	1.261	0.765	1.650	0.099	1.466	
Iter	Iter3_R	0.201	0.148	1.358	0.175	0.182	

Within the group analysis, Creativ questions 1 and 3 had higher significance, while Creativ Question 2 and Iter Questions 1-3 did not show any significance. What does this mean? Creative Question 2 probably doesn't correlate with the other two Creativ questions, and none of the Iter questions cluster or intercorrelate with one another effectively, although Iter Question 2 might be kept if needed ($p \cong 0.10$).

Design Framing *(Frame)* - Framing design problems is an important aspect of a design process.

Table 4 has previously shown that the Frame questions probe the degree to which students see some of the considerations needed within the framing of a design. While all Frame questions in Table 13 were significant in the pre-tests, some were less significant in the post-tests, including possibly Frame Questions 1 and 2. Since the questions were significant overall (but differed slightly in the amount of significance), probably all of the Frame questions cluster, are important in terms of measuring the latent variable, and intercorrelate with one another.

Table 13. Frame CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
Frame	Frame1	0.482	0.107	4.495	0.000	0.584	***
Frame	Frame2_R	0.360	0.092	3.920	0.000	0.380	***
Frame	Frame3	0.287	0.068	4.200	0.000	0.457	***
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
Frame	Frame1	0.348	0.111	3.147	0.002	0.453	**
Frame	Frame2_R	0.284	0.104	2.724	0.006	0.320	**
Frame	Frame3	0.499	0.139	3.581	0.000	0.804	***

Design Self-Efficacy *(DesSelfEff)* - These questions probe students' self-efficacy for designing.

Table 4 has previously that the DesSelfEff questions probe the degree to which students are confident in their ability to design. All DesSelfEff questions in

Table 14 were highly significant overall and therefore cluster, are important in terms of measuring the latent variable, and intercorrelate with one another.

Table 14. DesSelfEff CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
DesSelfEff	DesSelfEff1	0.693	0.044	15.689	0.000	0.831	***
DesSelfEff	DesSelfEff2	0.749	0.051	14.825	0.000	0.797	***
DesSelfEff	DesSelfEff3	0.741	0.053	13.974	0.000	0.768	***
DesSelfEff	DesSelfEff4	0.687	0.044	15.463	0.000	0.823	***
DesSelfEff	DesSelfEff5	0.545	0.057	9.538	0.000	0.571	***
DesSelfEff	DesSelfEff6	0.592	0.048	12.257	0.000	0.702	***
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
DesSelfEff	DesSelfEff1	0.708	0.057	12.444	0.000	0.854	***
DesSelfEff	DesSelfEff2	0.747	0.067	11.145	0.000	0.791	***
DesSelfEff	DesSelfEff3	0.813	0.074	10.947	0.000	0.784	***
DesSelfEff	DesSelfEff4	0.795	0.059	13.438	0.000	0.894	***
DesSelfEff	DesSelfEff5	0.480	0.068	7.022	0.000	0.556	***
DesSelfEff	DesSelfEff6	0.581	0.055	10.518	0.000	0.764	***

Intent to Persist *(IntPers)* - These questions probe students' intent to persist in an engineering career.

Table 4 has previously shown that the IntPers questions probe the degree to which students intend to persist in engineering careers. All IntPers questions in Table 15 were highly significant overall and therefore cluster, are important in terms of measuring the latent variable, and intercorrelate with one another.

Table 15. IntPers CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
IntPers	IntPers1	0.216	0.050	4.291	0.000	0.400	***
IntPers	IntPers2	-0.531	0.131	-4.065	0.000	-0.464	***
IntPers	IntPers3_R	-0.841	0.126	-6.652	0.000	-0.599	***
IntPers	IntPers4	0.562	0.149	3.780	0.000	0.427	***
IntPers	IntPers5	0.540	0.106	5.080	0.000	0.594	***
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
IntPers	IntPers1	0.480	0.068	7.081	0.000	0.652	***
IntPers	IntPers2	-0.500	0.127	-3.944	0.000	-0.398	***
IntPers	IntPers3_R	-0.827	0.143	-5.789	0.000	-0.543	***
IntPers	IntPers4	0.529	0.137	3.855	0.000	0.371	***
IntPers	IntPers5	0.750	0.096	7.830	0.000	0.737	***

Social Integration *(Social)* - These questions probe the social aspects of students' engineering major and career.

Table 4 has previously shown that the Social questions probe the degree to which students are involved in the social aspects of engineering education. Table 16 shows that social questions 1, 2, and 3 have larger significance in the post-tests than in the pre-tests but are significant in both. Questions 4 and 5 are less significant and could possibly be eliminated.

Table 16. Social CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
Social	Social1	0.591	0.299	1.976	0.048	0.348	*
Social	Social2	1.474	0.648	2.275	0.023	1.077	*
Social	Social3_R	-0.180	0.086	-2.101	0.036	-0.163	*

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Social	Social4	0.100	0.090	1.113	0.266	0.100	
Social	Social5_R	0.144	0.070	2.048	0.041	0.144	*
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
Social	Social1	1.042	0.187	5.569	0.000	0.617	***
Social	Social2	1.064	0.177	5.998	0.000	0.776	***
Social	Social3_R	-0.335	0.127	-2.636	0.008	-0.276	**
Social	Social4	0.297	0.108	2.739	0.006	0.269	**
Social	Social5_R	-0.215	0.116	-1.850	0.064	-0.197	

Degree Choice *(DegChoi)* - Is engineering a students' degree choice?

Table 4 has previously shown that the DegChoi questions probe the degree to which students are willing to persist in their degree choices. In Table 17, DegChoi Question 1 pretty clearly can be eliminated from the set as it has no significance in the pre-test and only mild significance in the post-test. Questions 2-6 are very significant in the pre-test, but only Questions 2-4 reiterate that significance in the post-test, which leaves room for the debate of Questions 5 and, particularly, 6, the latter of which holds no significance on the post-test results.

Table 17. DegChoi CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
DegChoi	DegChoi1_R	-0.007	0.107	-0.069	0.945	-0.006	
DegChoi	DegChoi2	0.966	0.123	7.885	0.000	0.748	***
DegChoi	DegChoi3	0.540	0.093	5.817	0.000	0.474	***
DegChoi	DegChoi4	0.413	0.065	6.324	0.000	0.471	***
DegChoi	DegChoi5	0.417	0.107	3.901	0.000	0.337	***

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
DegChoi	DegChoi6	0.329	0.096	3.431	0.001	0.286	***
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
DegChoi	DegChoi1_R	0.289	0.135	2.134	0.033	0.218	*
DegChoi	DegChoi2	1.035	0.156	6.652	0.000	0.736	***
DegChoi	DegChoi3	0.699	0.125	5.617	0.000	0.568	***
DegChoi	DegChoi4	0.591	0.115	5.121	0.000	0.522	***
DegChoi	DegChoi5	0.291	0.128	2.283	0.022	0.236	*
DegChoi	DegChoi6	0.236	0.126	1.867	0.062	0.192	

Design Challenge Motivation *(DesChalMot)* - These questions probe students' motivation for design challenges.

Table 4 has previously shown that the DesChalMot questions probe the degree to which students are motivated within the design challenges. All DesChalMot questions in Table 18 were highly significant overall and therefore cluster, are important in terms of measuring the latent variable, and intercorrelate with one another.

Table 18. DesChalMot CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
DesChalMot	DesChalMot1	0.494	0.051	9.632	0.000	0.706	***
DesChalMot	DesChalMot2	0.507	0.056	8.984	0.000	0.643	***
DesChalMot	DesChalMot3	0.461	0.052	8.891	0.000	0.634	***
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
DesChalMot	DesChalMot1	0.560	0.070	8.026	0.000	0.807	***
DesChalMot	DesChalMot2	0.422	0.070	5.995	0.000	0.549	***

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
DesChalMot	DesChalMot3	0.536	0.077	6.964	0.000	0.663	***

Professional Identity *(ProfIden)* - These questions probe students' ability to identify with engineering professionals.

Table 4 has previously shown that the ProfIden questions probe the degree to which students identify with engineering careers. In Table 19, ProfIden questions 2, 3, and 5-7 were highly significant overall and therefore cluster, are important in terms of measuring the latent variable, and intercorrelate with one another. Question 4 was less significant in the Post-test analysis than in the Pre-test analysis, but was significant in both, so it could be an optional elimination from the survey.

Table 19. ProfIden CFA Individual Question Analyses by Pre-Test vs. Post-Test

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
ProfIden	ProfIden2	0.353	0.065	5.425	0.000	0.374	***
ProfIden	ProfIden3	0.515	0.071	7.297	0.000	0.491	***
ProfIden	ProfIden4	0.461	0.109	4.236	0.000	0.288	***
ProfIden	ProfIden5	0.407	0.041	9.863	0.000	0.616	***
ProfIden	ProfIden6	0.718	0.051	14.020	0.000	0.824	***
ProfIden	ProfIden7	0.572	0.045	12.609	0.000	0.745	***
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)							
ProfIden	ProfIden2	0.263	0.077	3.418	0.001	0.313	***
ProfIden	ProfIden3	0.397	0.093	4.295	0.000	0.388	***
ProfIden	ProfIden4	0.347	0.145	2.395	0.017	0.223	*
ProfIden	ProfIden5	0.453	0.077	5.892	0.000	0.525	***
ProfIden	ProfIden6	0.744	0.075	9.978	0.000	0.819	***

Latent Factor	Indicator	B	SE	Z	p-value	Beta	sig
ProfIden	ProfIden7	0.617	0.066	9.343	0.000	0.770	***

The analysis of the fit indices for the confirmatory factor analysis is shown in Table 20. In Table 20, the model baseline fit p-value, the CFI, the RMSEA and the SRMR show that the model fit is good in both the pre- and post-test data for the latent variables MeetNeeds, Creativ, Frame, and DesMotChal. The model baseline fit p-value, the CFI, and the SRMR show that the model fit is relatively good in both the pre- and post-test data for DesSelfEff. The model baseline fit p-value, the CFI, and the SRMR show that the model fit is okay in both the pre- and post-test data for IllStruc and ProfIden, The CFI, the RMSEA and the SRMR show that the model fit is relatively bad in both the pre- and post-test data for the latent variables IntPers, Social, and DegChoi, which is likely due to low correlation values between the questions, and may need further analysis of the covariances between the latent factors.

Table 20. Fit Indices for the Confirmatory Factor Analysis. Variable name, Model χ^2 test, CFI, RMSEA, and χ^2 comparing the orthogonal and proposed models are listed.

Variable	Model χ^2 Baseline Fit p-value	Degrees of Freedom	CFI	RMSEA	SRMR
Pre-Test					
MeetNeeds	<0.005	3	1.000	0.000	0.000
IllStruc	<0.005	10	0.833	0.126	0.059
Iter	Results were not normal; therefore, fit could not be measured				
Creativ	<0.005	3	1.000	0.000	0.000
Frame	<0.005	3	1.000	0.000	0.000
DesSelfEff	<0.005	15	0.929	0.159	0.041
IntPers	<0.005	10	0.664	0.214	0.089
Social	<0.005	10	0.783	0.105	0.058
DegChoi	<0.005	15	0.611	0.166	0.089

DesChalMot	<0.005	3	1.000	0.000	0.000
ProfIden	<0.005	15	0.833	0.164	0.079
Post-Test					
MeetNeeds	<0.005	3	1.000	0.000	0.000
IllStruc	<0.005	10	0.837	0.128	0.060
Iter	Results were not normal; therefore, fit could not be measured				
Creativ	<0.005	3	1.000	0.000	0.000
Frame	<0.005	3	1.000	0.000	0.000
DesSelfEff	<0.005	15	0.960	0.124	0.031
IntPers	<0.005	10	0.804	0.179	0.072
Social	<0.005	10	0.766	0.147	0.075
DegChoi	<0.005	15	0.658	0.160	0.083
DesChalMot	<0.005	3	1.000	0.000	0.000
ProfIden	<0.005	15	0.849	0.146	0.074

Principal Component Analysis (PCA)

The same datasets were used for the principal component analysis as for the confirmatory factor analyses. Principal component analysis and confirmatory factor analysis were required to answer research questions 1-3. The first ten principal components are shown in Table 21, which lists the standard deviation of each component, the proportion of variance explained by that specific component, and the overall variance (cumulative proportion) tracked as each component is factored into the model. In total, all ten principal components explain close to 60% of the total variance in each of the analyses – Pre and Post-test.

Table 21. Principal Component Analysis for both Pre-tests and Post-tests, including the standard deviation and proportional variance of each of the first ten principal components.

Component Number	Standard Deviation	Proportion of Variance	Cumulative Proportion
Pre-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)			
1	2.716	0.151	0.151
2	1.898	0.074	0.224
3	1.749	0.062	0.286
4	1.725	0.061	0.347
5	1.584	0.051	0.398

6	1.469	0.044	0.442
7	1.355	0.037	0.480
8	1.320	0.036	0.516
9	1.254	0.032	0.548
10	1.235	0.031	0.579
Post-Tests (Fall 2015, Fall 2016, Fall 2017, Fall 2018)			
1	2.845	0.155	0.155
2	2.074	0.0827	0.238
3	1.865	0.0668	0.305
4	1.769	0.0601	0.365
5	1.593	0.0488	0.414
6	1.544	0.0458	0.460
7	1.434	0.0395	0.499
8	1.387	0.0370	0.536
9	1.368	0.0360	0.572
10	1.312	0.0330	0.605

Table 21 shows that the data seems spherical – PC1 and PC2 (Principal components 1 and 2, respectively) are typically larger and account for a larger portion of the variance. This set of principal components just regularly increases by similar increments, which means that principal component analysis is not very helpful for this analysis.

A scree plot shows the proportion of variance explained by each component, which corresponds that principal component's eigenvalue divided by all of the eigenvalues. The scree plots of the pre-test and post-test (Figure 3) are shown below. These plots also give us a quick view of the cumulative variance over the first ten principal components.

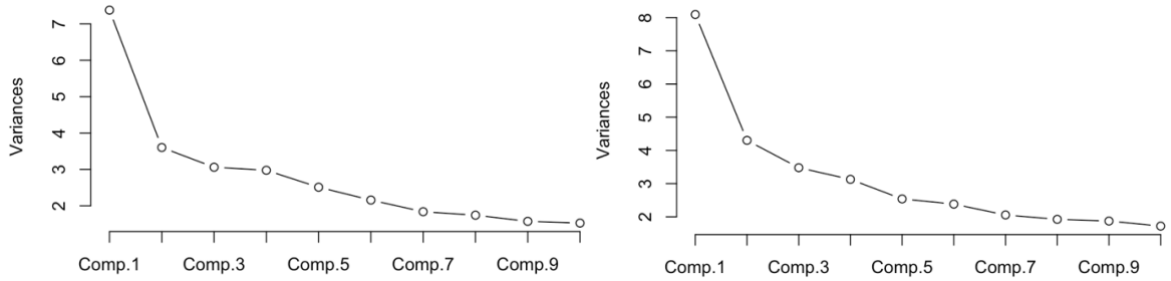

Figure 3. Pre-Test Scree Line Plot on the Left, Post-Test Scree Line Plot on the Right

The biplots shown in Figures 4 (pre-test) and 6 (post-test) represent some of the most helpful graphical representations of a large multivariate dataset. The variables (the survey questions) are shown as red arrows and the black numbers correspond to study IDs (or the number of students). The bottom axis represents the scores for principal component one (PC1 or Comp. 1) and the left axis represents the scores for principal component two (PC2 or Comp. 2); these axes are used to evaluate the numbers. The top axis represents the loadings on Comp. 1 and the right axis represents the loadings on Comp. 2; these axes are used to evaluate the arrows, specifically how strongly each factor (i.e. vector) influences the principal components. Those vectors that are overlapping indicate the factors that are highly correlated. Those vectors that are more parallel to the x-axis (more horizontal) correlate strongly with Comp. 1 and those vectors that are more perpendicular to the x-axis (more vertical) correlate strongly with Comp. 2. We can see that the pre-test biplot (Figure 4) and the post-test biplot (Figure 6) show very similar results, including the similarity in the loadings on the variables. The correlations between the overlapping vectors and the correlations to the principal components is similar between the pre- and the post-test PCAs.

Figures 5 (pre-test) and 7 (post-test) are graphs that visualize the variables more closely, showing the contributions of the variables to principal components 1 (shown here as Dim1) and 2 (shown here as Dim2). The contributions are calculated as:

*contrib = (Contribution of the variable on PC1 * Eigenvalue of PC1) + (Contribution of the variable on PC2 * Eigenvalue of PC2)*

The highest contributions to PC1 (Dim1) and PC2 (Dim2) in the pre-tests shown in Figure 5 are from DesSelfEff Questions 1, 2, 4, and 6. The highest contributions to PC1 (Dim1) and PC2 (Dim2) in the post-tests shown in Figure 7 are from DesSelfEff Questions 1, 2, and 4.

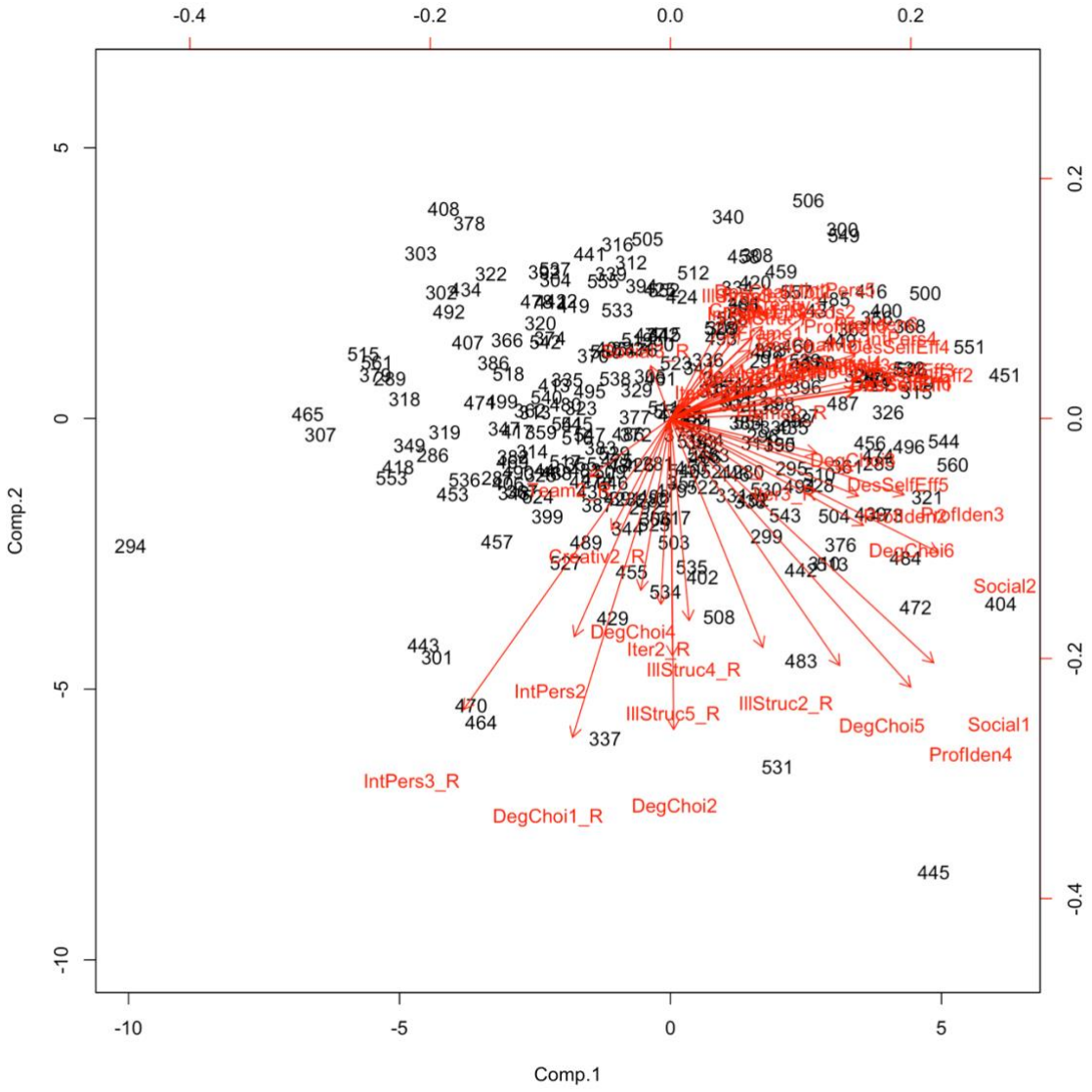

Figure 4. Biplot of first two principal components of the Pre-test PCA analysis. Principal component 1 (Comp. 1 or PC1) explains about 15.1% of the variance in the analysis, and principal component 2 (Comp 2 or PC2) explains about 7.4% of the variance in the analysis.

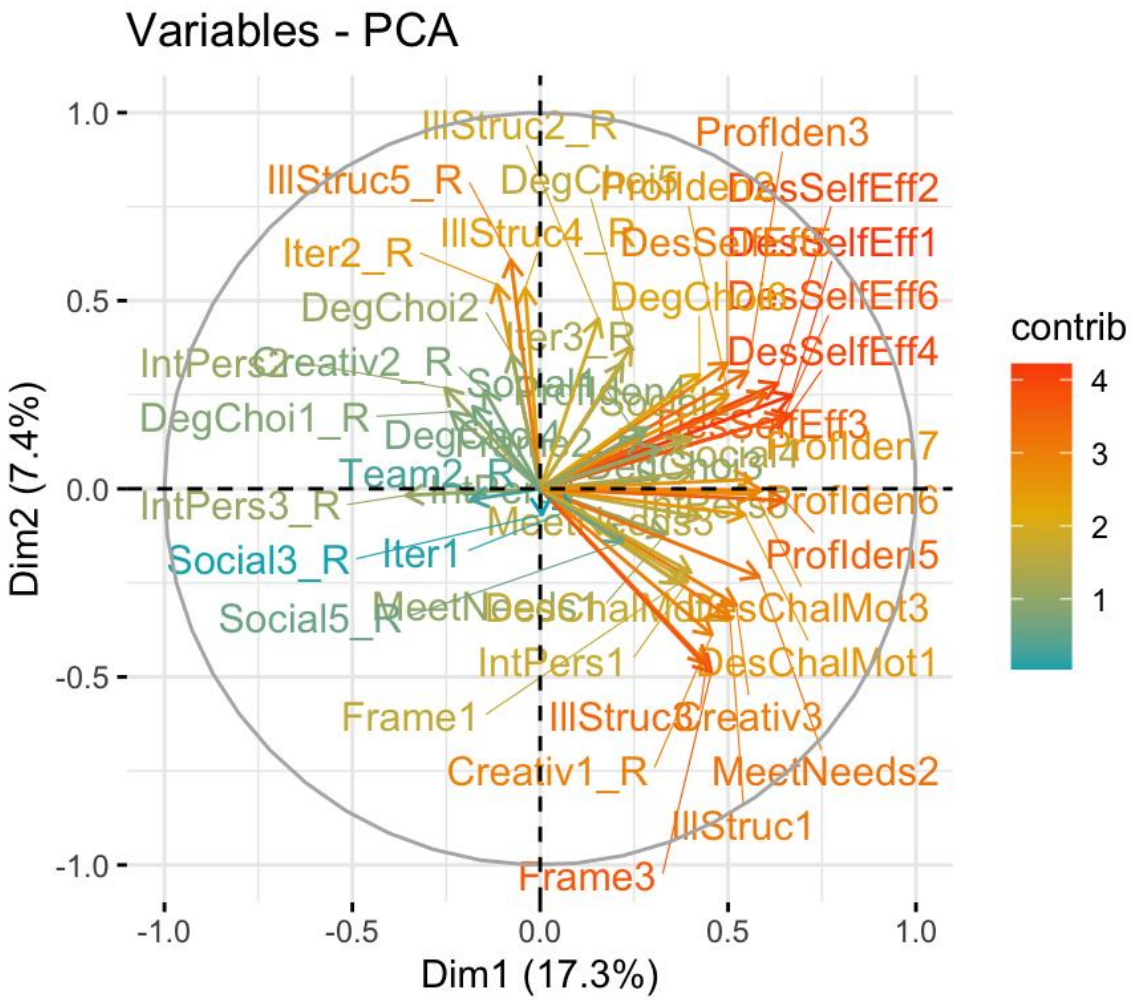

Figure 5. The PCA Variable graph for the pre-tests, showing each variables' relative contribution to PC1 (Dim 1) and PC2 (Dim 2). The highest contributions to PC1 (Dim1) and PC2 (Dim2) in the pre-tests are from DesSelfEff Questions 1, 2, 4, and 6.

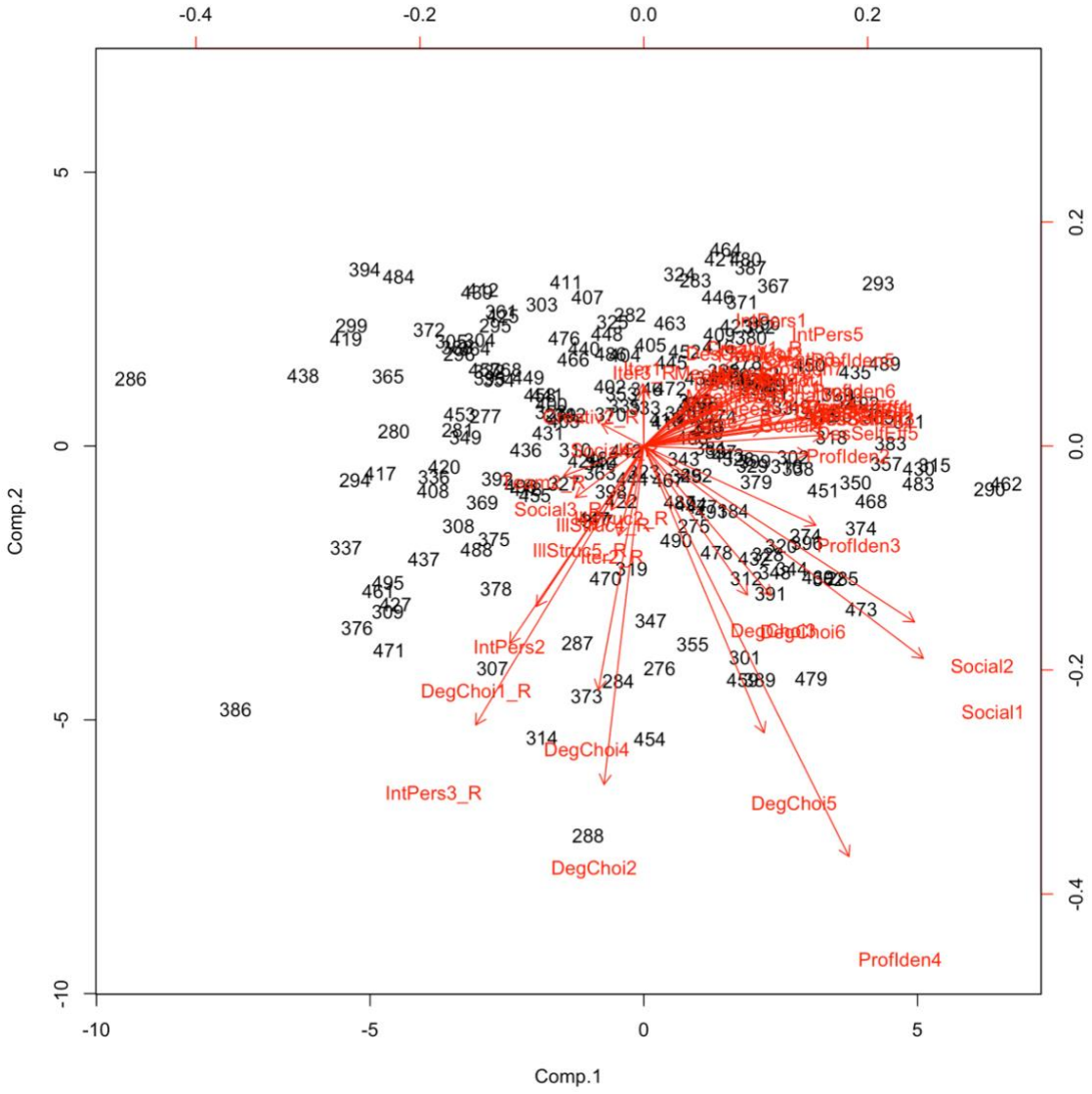

Figure 6. Biplot of the first two principal components of the Post-test PCA analysis. Principal component 1 (Comp. 1 or PC1) explains about 15.5% of the variance in the analysis, and principal component 2 (Comp 2 or PC2) explains about 8.3% of the variance in the analysis.

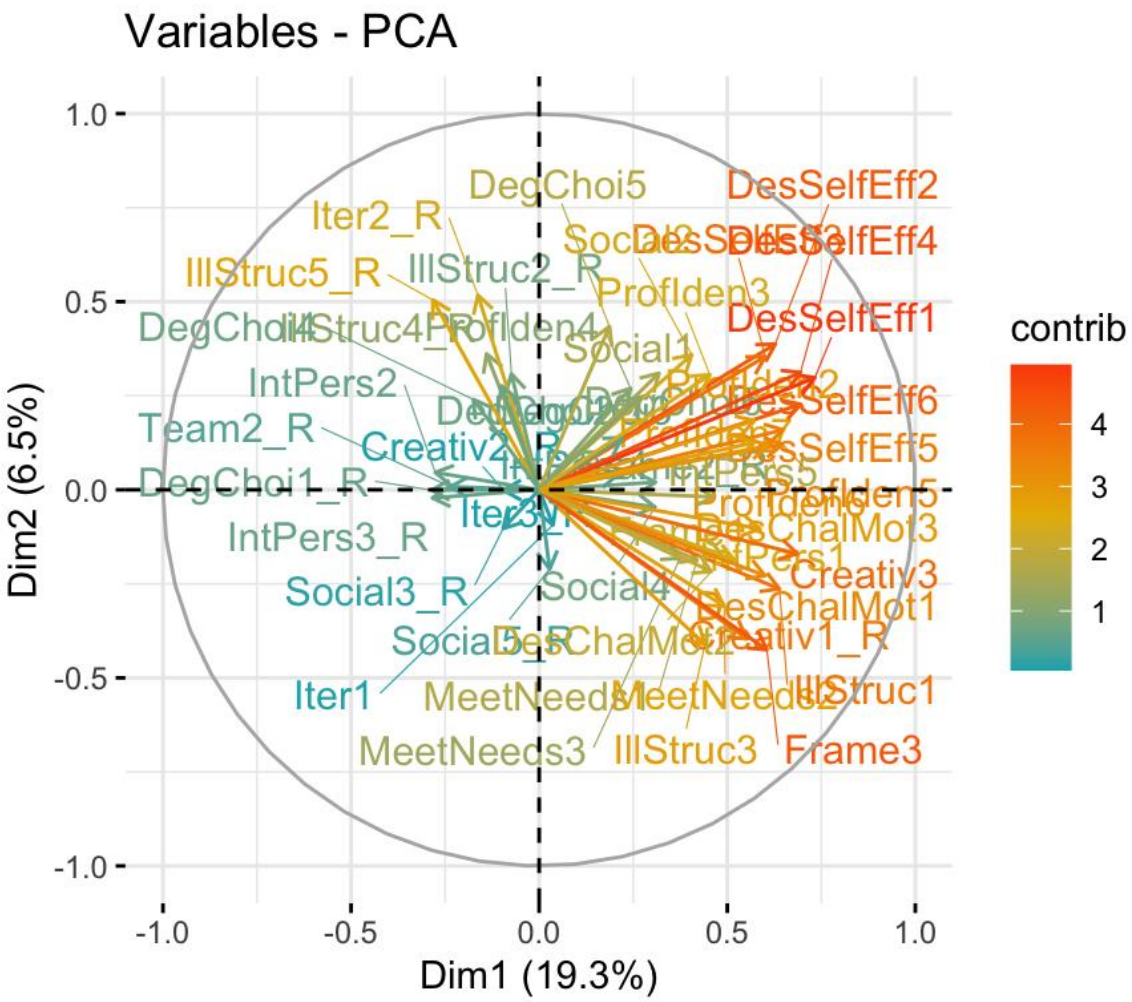

Figure 7. The PCA Variable graph for the post-tests, showing each variables' relative contribution to PC1 (Dim 1) and PC2 (Dim 2). The highest contributions to PC1 (Dim1) and PC2 (Dim2) in the post-tests are from DesSelfEff Questions 1, 2, and 4.

Cluster Analysis

Cluster analysis was used to answer research questions 2 and 3. Table 22 describes the clusters obtained from kmeans using the sum of squares distance, but it is difficult to visualize this distance (as well as the clusters) without a graphical representation.

Table 22. K means cluster analysis size and sum of squares distance between groups.

Cluster Number	Cluster Size Pre-Test	Within Cluster Sum of Squares by Cluster (Pre-Test)	Cluster Size Post-Test	Within Cluster Sum of Squares by Cluster (Post-Test)
1	26	965	14	457
2	30	1068	27	774
3	21	676	24	737
4	15	571	22	684
5	14	523	12	418
6	25	838	24	737
7	18	728	12	473
8	23	697	14	627
9	28	955	20	745
10	11	334	14	527
11	4	186	14	532
12	42	1387	16	613

Figures 8 and 9 show the kmeans cluster plot for the pre-test data and the post-test data, respectively. In each of these figures, we see the pre-test or post-test kmeans model data plotted against the first two discriminant functions using a cluster plot. In this plot, the distance between clusters should be large with minimal overlap. Ideally, the numbers of each cluster would be closely grouped together, and the clusters would be as far apart as possible.

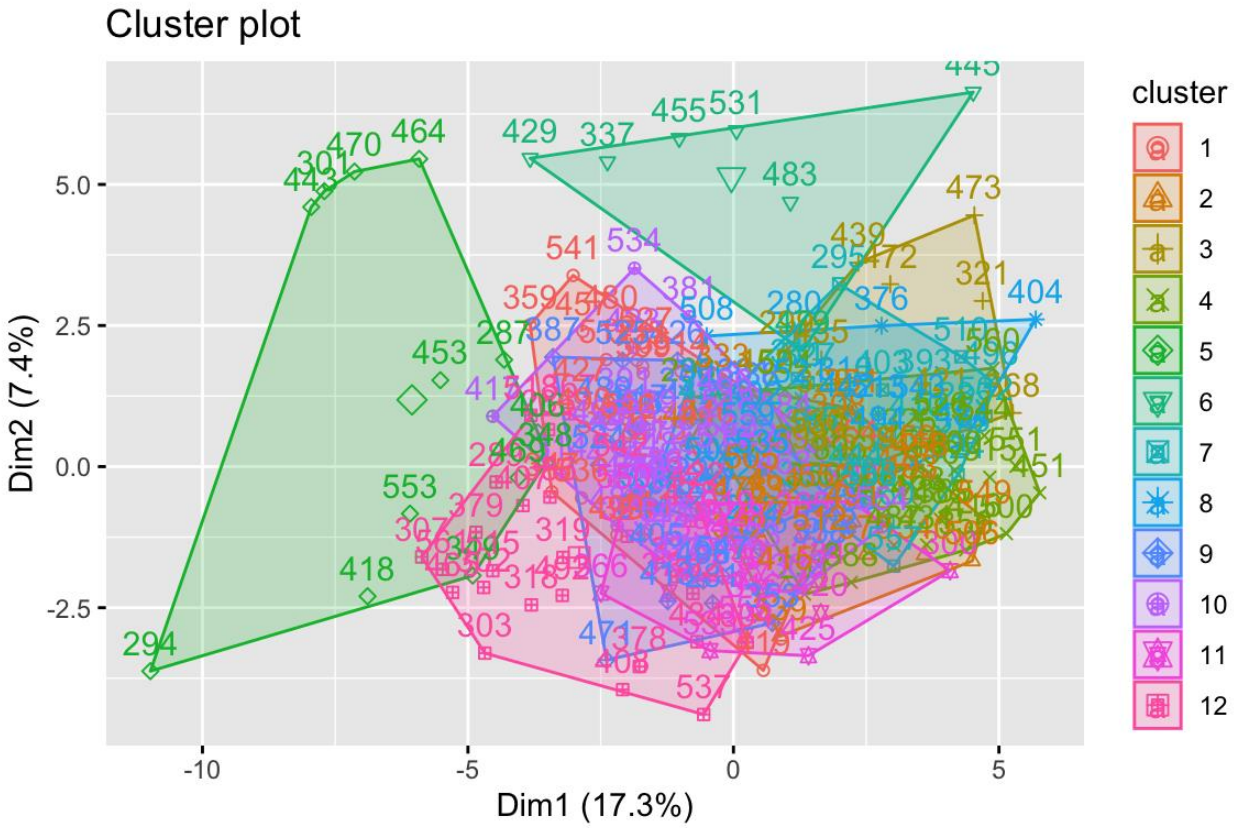

Figure 8. Kmeans cluster data analysis with twelve groups performed on the pre-test data. The plot is grouped in clusters in which the kmeans model is plotted against the first two discriminant functions (or PC1 and PC2, respectively). We can see from the plot that the kmeans cluster analysis needs further analysis as the current clusters significantly overlap.

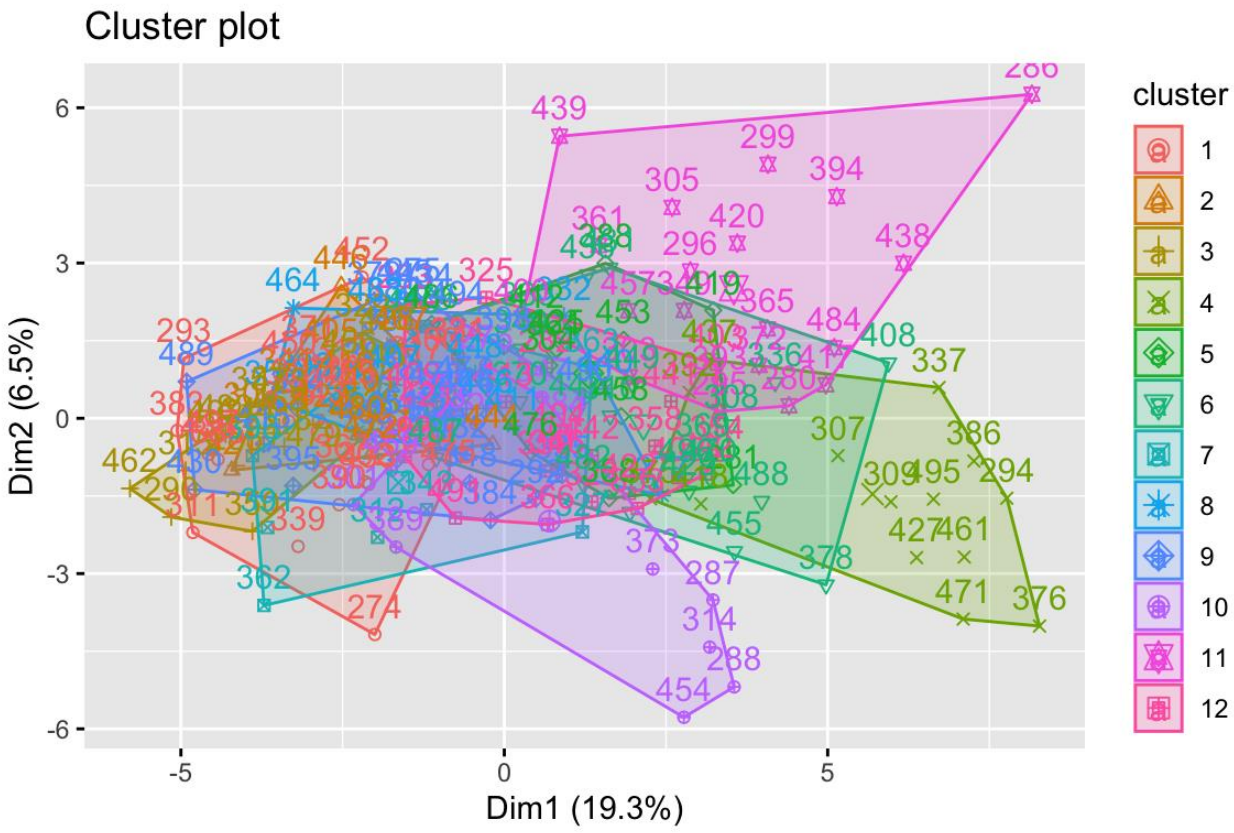

Figure 9. Kmeans cluster data analysis with twelve groups performed on the post-test data. The plot is grouped in clusters in which the kmeans model is plotted against PC1 (Dim1) and PC2 (Dim2). We can see from the plot that the kmeans cluster analysis needs further analysis as the current clusters significantly overlap.

In both cluster plots, we can see that there is significant overlap between clusters and that there is not tight grouping around the centroid in the clusters on either plot. Twelve clusters is not an ideal number of clusters.

The elbow method, which is used to determine the ideal numbers of clusters, was employed to find the optimal number of clusters in k-means. Twelve clusters were seen as the least ideal, but there is no distinct elbow - a point where the scree plot transitions from a steep decline to a flatter region. Figure 10 shows the scree plots of the total within clusters sum of squares (i.e., distance between the clusters)

and the number of clusters for both the pre-test and post-test data. Two clusters seem to have the largest between-cluster distance in both plots, but as there is no clear elbow here, two or three clusters may yield more distinct clusters.

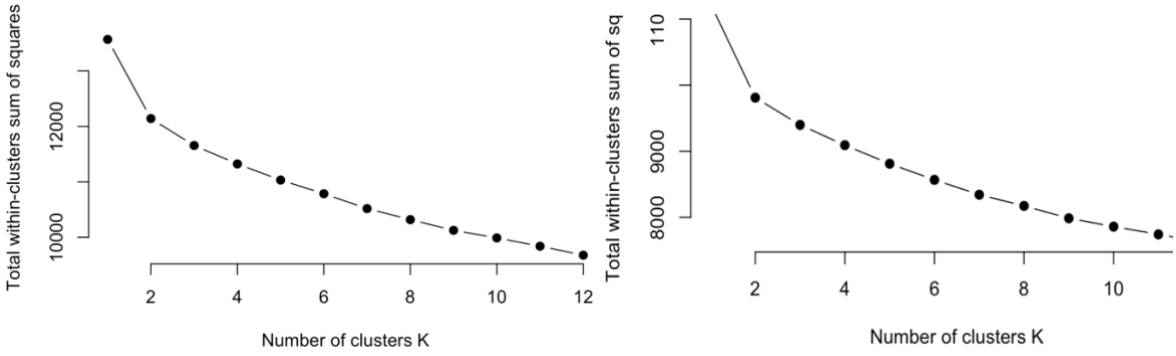

Figure 10. Scree plot of elbow method to find an optimal number of clusters. The pre-test data scree plot is on the left and the post-test data scree plot is on the right.

The kmeans analysis was run again, systematically increasing the k value from two to twelve to determine the optimal number of clusters for the kmeans analysis. I used *kmeansrun*, an R package that runs kmeans but initializes the algorithm several times with random points from the dataset and then estimates the optimal number of clusters by either the Calinski Harabasz index or the average silhouette width. Figure 11 shows a plot of the cluster output using the average silhouette width estimation on the left and the Calinski Harabasz index estimation on the right.

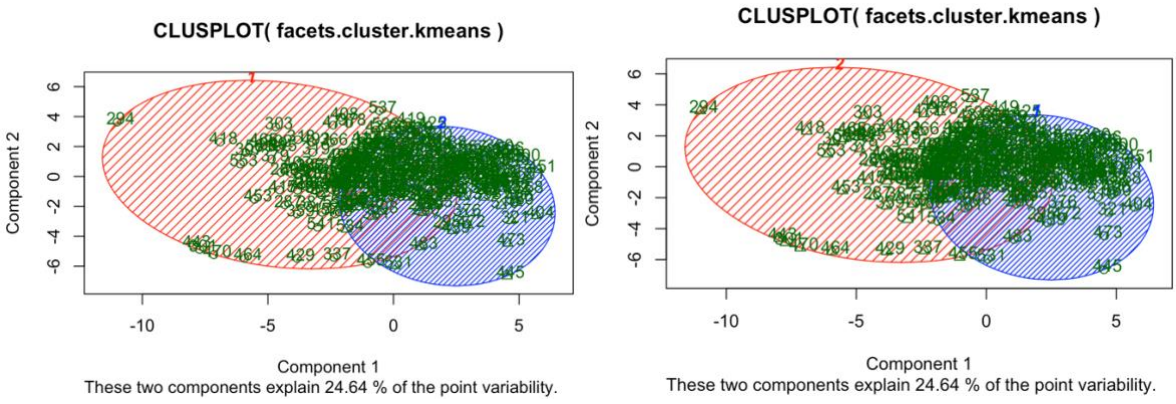

Figure 11. Optimum clustering for kmeans shows two clusters in total. The cluster plot with average silhouette width estimation is on the left and the cluster plot with Calinski Harabasz index estimation is on the right. The two clustering estimations look virtually the same with minimal differences.

I plotted the BIC (Bayes Information Criteria) for all the models with profiles ranging between 1 and 9 in Figure 12, but it is unclear as to what model is best because several of the models overlap, so the BIC chooses the top three models. The partitions are obtained from agglomerative hierarchical clustering.

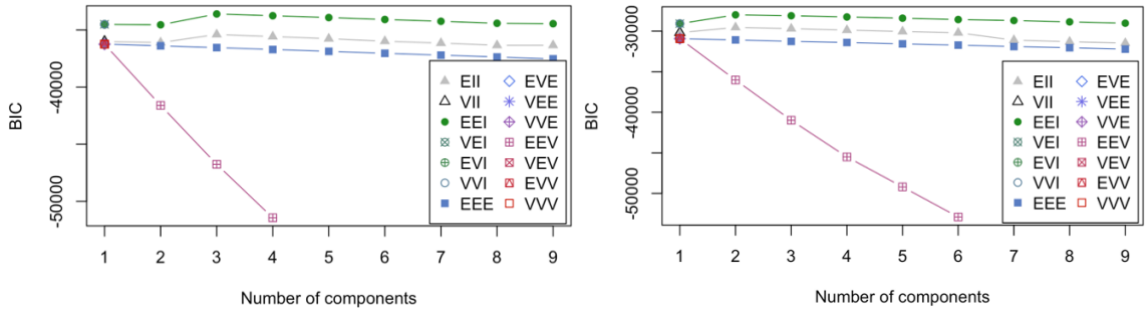

Figure 12. Plotted Bayes Information Criteria for the first nine covariance models for both the Pre-test and Post-test data. This kind of plot summarizes the top three models and the BIC traces for all the models considered. It is unclear as to which model is best since they mostly overlap.

The Bayes Information Criteria (BIC) values for the best models are listed in Table 23. Highlighted in green for each set of data is the best model and the BIC differences column shows the difference between the BIC of the model listed and the best model.

Table 23. BIC values for the three best models for clustering using *mclust* for both the pre-test and post-test data. Those models highlighted in green are considered the best overall. EEI, 3, which is the third expectation-maximization model using EEI, described on Table 21 as a diagonal distribution with equal volume and shape, is the best model for the pre-test data and is the second best for the post-test data. Therefore, it is probably the best model overall as well.

	BIC	BIC differences

Pre-Test Data		
EEI, 3	-33603.01	0.00
EEI, 4	-33757.80	-154.79
EEI, 5	-33917.79	-314.78
Post-Test Data		
EEI, 2	-27995.34	0.00
EEI, 3	-28104.32	-108.98
EEI, 4	-28254.42	-259.08

The best model for the pre-test has 3 components, is EEI (a diagonal distribution with equal volume and shape), a log-likelihood of -16252.15, and a BIC of -33603.01 with 198 degrees of freedom. The best model for the post-test has 2 components, is EEI (a diagonal distribution with equal volume and shape), a log-likelihood of -13600.93, and a BIC of -27995.34 with 148 degrees of freedom. Since the second-best model for the post-test is the same as the best for the pre-test, it's most likely the best model overall.

Grouped cluster analysis was run with *mclust* in R. The variables needed to be grouped due to the large number of variables within the datasets. The groups, in terms of latent variables described on Table 3, were: 1) MeetNeeds, IllStruc, Iter and Creativ in the first *mclust* analysis, which correspond to the first row of density plots in Figure 13; 2) Frame, DesSelfEff, and IntPers in the second *mclust* analysis, which correspond to the second row of density plots in Figure 13; and 3) Social, DegChoi, DesChalMot and ProfIden in the third *mclust* analysis, which correspond to the third row of density plots in Figure 13. In the Pre-test data (the plots on the right of Figure 13), IllStruc3, IntPers1, and ProfIden5 showed the least overlap in the density plots. While the Post-test data (the plots on the left of Figure 13) shows

significantly less overlap overall in the first two plots, IllStruc3 and IntPers1 showed less density overlap than the rest of the density plots. Within the last density Post-test plot, while ProfIden5 has considerable overlap, DesChalMot1 has significantly less density overlap.

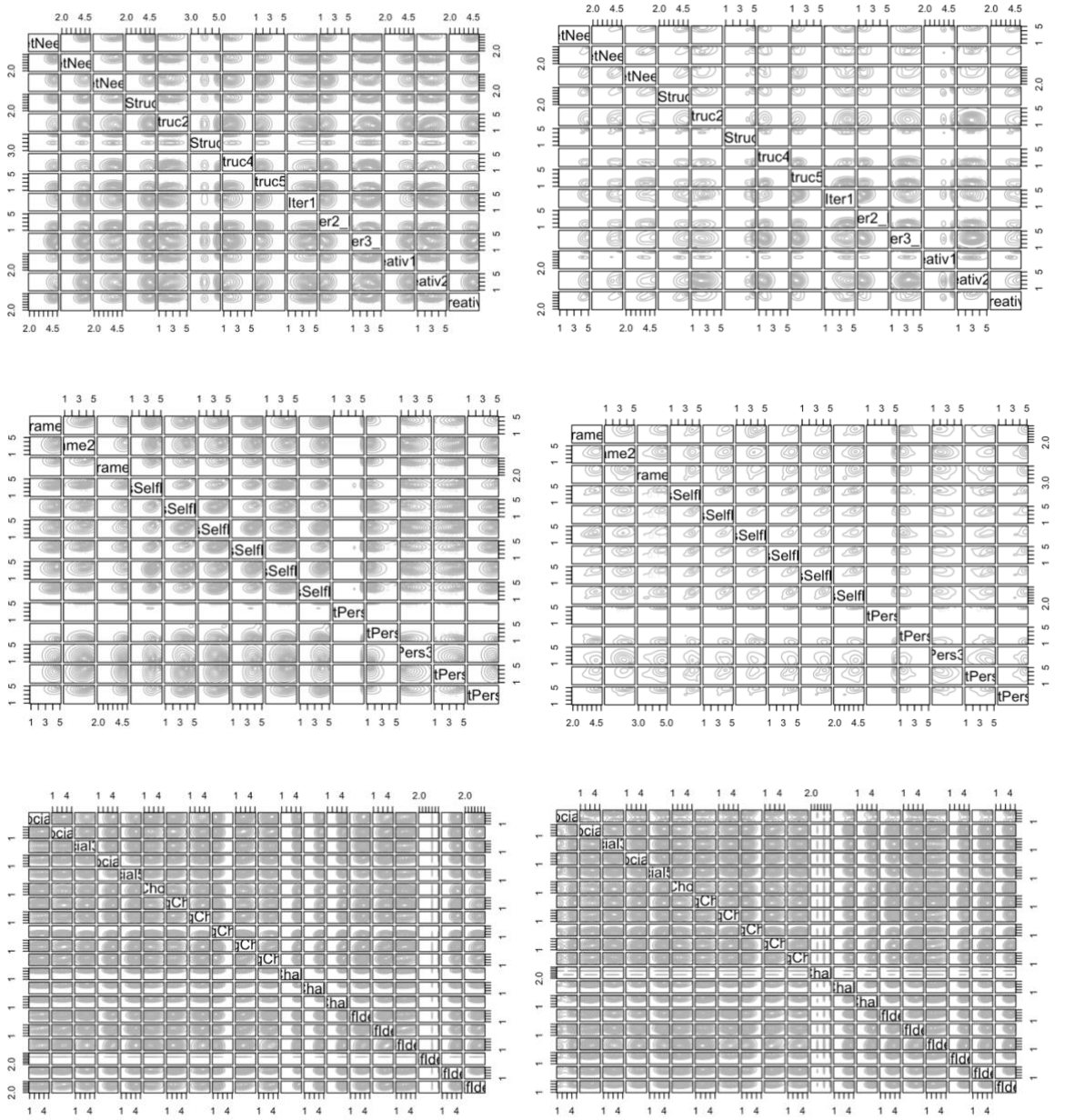

Figure 13. Density plots from *mclust* analysis in R. On the left of this figure, the grouped cluster analysis for the pre-tests shows less density overlap for IllStruct3, IntPers1, and ProfIden5. On the right of this figure, the grouped

cluster analysis for the post-tests shows less density overlap for IllStruct3, IntPers1, and DesMotChal1.

We can validate the cluster analysis by calculating cluster validation statistics using a comparison between the kmeans cluster analysis and the normal finite mixed method used in *mclust.* The comparison, which is based on distance-based statistics, yields information about distance between the clusters using the Euclidean distance:

$$d = \sqrt{\sum_{i=1}^{N} (cluster\ matrix_{kmeans} - \ cluster\ matrix_{mclust})^2}$$

The larger the distance between the clusters, the better the clustering analysis model used (kmeans vs. *mclust*). Preliminary analysis shows larger distances in the *mclust* analysis since the kmeans clusters overlap extensively, so therefore, *mclust* might be considered a better method.

The Survey is Shortened and Reanalyzed

Based on the CFA and PCA results, I advised eliminating several questions from the original survey design (Table 24). These questions were considered problematic because: 1. they did not cluster with the other questions as expected and/or 2. they did not vary over time or 3. they varied too wildly over time.

Table 24. Survey questions eliminated from the original survey for Fall 2019. These questions were eliminated based on Dr. Svihla's extensive knowledge of the survey as well as the CFA and PCA analyses.

Variable name	Question text	Question Response
IllStruc1_YEAR MONTH	In design, the problem and the solution co-evolve, where an advance in the solution	5-point Likert scale from strongly agree (1) to strongly disagree (5)

	leads to a new understanding of the problem.	
IllStruc4_YEARMONTH_R	Designers of equal skill and experience should come to the same design solution given the same initial design problem	5-point Likert scale (reversed) from strongly disagree (1) to strongly agree (5)
Iter1_YEARMONTH	Design is iteration	5-point Likert scale from strongly agree (1) to strongly disagree (5)
Iter3_YEARMONTH_R	Design is a goal-oriented, constrained activity	5-point Likert scale (reversed) from strongly disagree (1) to strongly agree (5)
Creativ1_YEARMONTH_R	Expert designers typically consider many possible ideas which leads to better solutions	5-point Likert scale (reversed) from strongly disagree (1) to strongly agree (5)
Creativ2_YEARMONTH_R	Constraints typically hinder creative design	5-point Likert scale (reversed) from strongly disagree (1) to strongly agree (5)
Creativ3_YEARMONTH	Creativity is integral to design. Every design project involves creativity.	5-point Likert scale from strongly agree (1) to strongly disagree (5)
DesSelfEff2_YEARMONTH	I am confident I could select the best possible design for an authentic engineering design problem	5-point Likert scale from strongly agree (1) to strongly disagree (5)
Social4_YEARMONTH	The faculty and staff make engineering feel like a welcoming place for me	5-point Likert scale from strongly agree (1) to strongly disagree (5)
Social5_YEARMONTH_R	It is very important to me to be involved in non-engineering activities, such as hobbies, civic or church organizations, campus publications, student government, social fraternity or sorority, sports, etc.	5-point Likert scale (reversed) from strongly disagree (1) to strongly agree (5)

I ran an initial analysis of this same dataset with the questions from Table 24 eliminated to try to get a sense of how the statistics would change with the shortened survey data. I found that for those confirmatory factor analysis (CFA) results that were problematic and that had questions removed, the CFA results from the modified dataset vastly improved in terms of the fit of the model. These results are shown in Table 25, which describes the latent variable name, the comparative fit index (CFI), the Root-Mean Square Error of Approximation (RMSEA), and the

significance of the individual indicators, which are assumed to be at least significant at a 0.01 alpha value unless otherwise stated.

Table 25. The CFA fit values and question significance results for the shortened survey.

Latent Variable	Test	CFI	RMSEA	Significance of Indicators
MeetsNeeds	Pre and post	1.000	0.000	All questions were significant at the 0.01 or 0.001 alpha levels.
IllStruc	Pre and post	1.000	0.000	All questions were significant at the 0.01 or 0.001 alpha levels.
Iter	Pre and post	1.000	0.000	The remaining question was significant at the 0.001 alpha level.
Frame	Pre and post	1.000	0.000	All questions were significant at the 0.01 or 0.001 alpha levels.
DesSelfEff	Pre	0.903	0.215	All questions were significant at the 0.01 or 0.001 alpha levels.
	Post	0.951	0.161	
IntPers	Pre	0.664	0.214	All questions were significant at the 0.01 or 0.001 alpha levels.
	Post	0.804	0.179	
Social	Pre and Post	1.000	0.000	Questions 1 and 2 had significance at the 0.001 alpha level. Question 3 was significant at the 0.05 alpha level in both pre- and post-tests.
DegChoi	Pre	0.611	0.166	All questions were significant at the 0.01 or 0.001 alpha levels except question 1, which was not significant at any level in both the pre- and post-tests.
	Post	0.658	0.160	
DesChalMot	Pre and Post	1.000	0.000	All questions were significant at the 0.01 or 0.001 alpha levels.
ProfIden	Pre	0.833	0.164	All questions were significant at the 0.01 or 0.001 alpha levels.
	Post	0.849	0.146	

Upon deleting the questions in Table 24 from the original dataset, IllStruc, Iter, and Social had both model fit values that vastly improved and questions that were significant at higher levels. The model fits for IntPers and DegChoi, which both had no questions deleted, are still questionable, and if the questionable fit is due to a lack of correlation amongst the values, perhaps these questions need to be

revisited as well. Overall, deleting the questions in Table 24 from the survey seems to have improved the statistical analysis, in that the model fit of the CFA is better.

The first ten principal components in the PCA run on the shortened survey accounted for 63.95% and 65.16% of the variance in the pre- and post-test, respectively. The Bayes Information Criteria from the *mclust* analysis shows that the model EEI, 3 is best on the pre-test but VEI, 2 is the best on the post-test. EEI, 3 ranked third-best on the post-test. We can see from Figure 14 that the cluster analysis has not vastly improved even with the questions from Table 24 deleted from the survey.

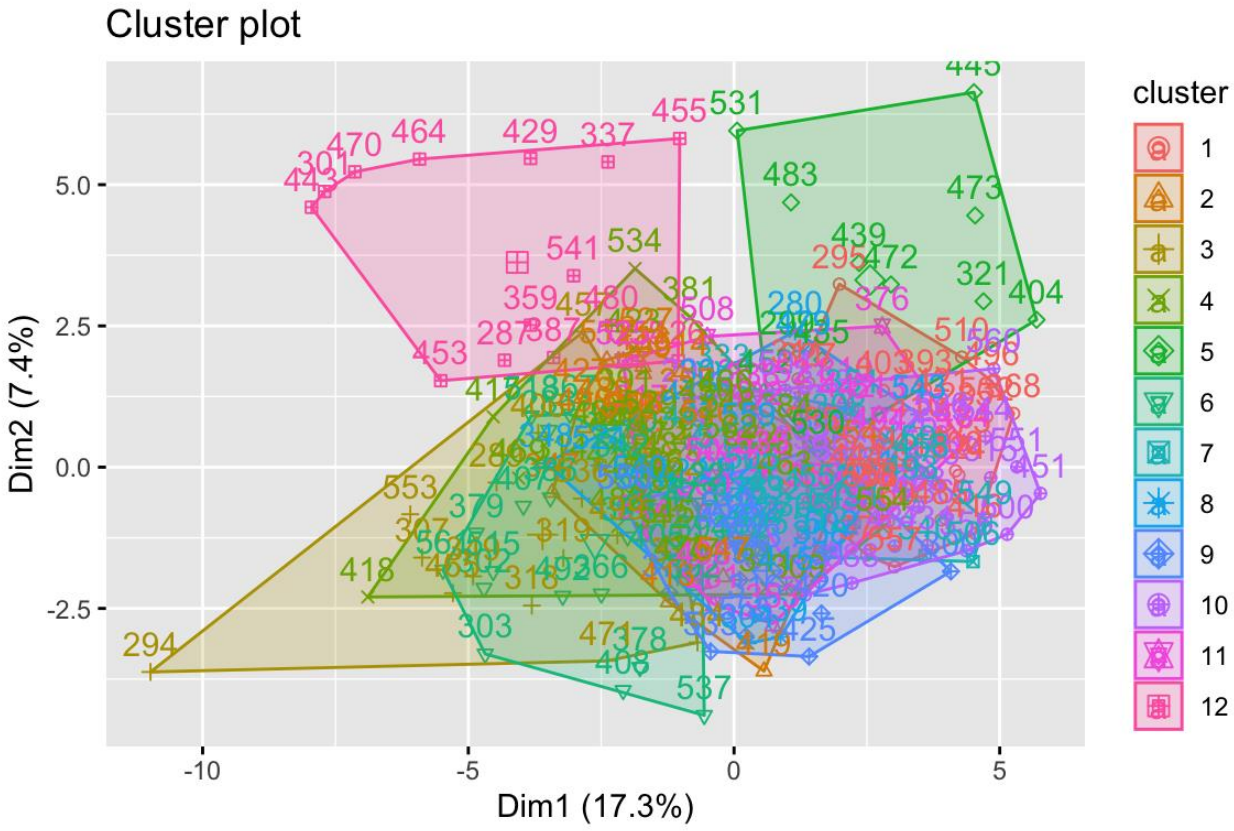

Figure 14. Kmeans cluster data analysis with twelve groups performed on the pre-test data. The graph shows the kmeans model plotted against the first two principal components (Dim1 and Dim2). We can see from the plot

that the kmeans cluster analysis needs further analysis as the current clusters significantly overlap.

Linear Regression Analysis and ANOVAs (Pilot and Shortened Dataset)

A pilot linear regression analysis was performed on the dataset for the longer survey. We will refer to this analysis as the full additive model. The response variable was IntPers or the student's intent to persist within an engineering degree. No interactions were included since we only include interactions in educational research if we see a theoretical or empirical need to incorporate them. What we can see from Table 26 below is that all of the main effects contribute significantly to the full additive model except for DesSelfEffAVG (or the average value across all questions referring to Design Self Efficacy) and Semester 2, which is a factor variable referring to the pre-test for Fall 2016.

Table 26. The statistical analysis of the linear regression is shown for the full model. The columns correspond to β_0(estimates), which correspond to the loadings on each variable or specific slope according to that variable; SE, or the standard error; t value, or the test run to determine the significance of each main effect to the overall model; and the p-value, the probability that we will obtain test results from the t-test that are at least as extreme as those observed.

Variable Name	β (Estimate)	SE	t value	p-value
Intercept (β_0)	2.254	0.142	15.838	$< 2 \times 10^{-16}$
DesSelfEffAVG	0.0373	0.0353	1.056	0.291
SocialAVG	0.204	0.0364	5.604	3.19×10^{-8}
ProfIdenAVG	0.223	0.0415	5.373	1.10×10^{-7}
Semester 1	-0.260	0.0760	-3.418	0.000674
Semester 2	-0.0597	0.0694	-0.859	0.391
Semester 3	-0.284	0.0710	-3.996	7.24×10^{-5}
Semester 4	-0.528	0.116	-4.572	5.85×10^{-6}
Semester 5	-0.326	0.101	-3.226	0.00132
Semester 6	-0.532	0.128	-4.142	3.94×10^{-5}

The ANOVA statistics included in Table 27 confirm that all main effects contribute significantly to the variation in the full model. Therefore, we might infer from Tables 26 and 27 that even though three of the four main effects is significant in the Regression Table and all four main effects are significant in ANOVA table (which makes sense because ANOVA treats the variables as factors and Likert scales are ordinal factors), further analysis is needed as we may need to transform variables within the linear regression to account for DesSelfEffAVG or we may need to revisit each individual question in DesSelfEffAVG within the linear regression to account for individual variability.

Table 27. ANOVA Type I Statistics. The alpha value for the factors in this dataset is set at 0.05 and any p-value that is below that alpha value contributes significantly to the variation in the model. Looking, then, at the p-value column, we can see that all of the main effects contribute most significantly to the variation in the model.

ANOVA (Analysis of Variance)					
Source of Variation	**df**	**Sum of Squares (SS)**	**Mean Squares (MS = SS/df)**	**F value**	**p-value**
DesSelfEffAVG	1	14.835	14.835	47.766	1.22×10^{-11}
SocialAVG	1	10.414	10.414	33.529	1.13×10^{-8}
ProfIdenAVG	1	14.599	14.599	47.003	1.75×10^{-11}
Semester	6	14.048	2.341	7.528	8.24×10^{-8}
Residuals	605	187.905	0.311		

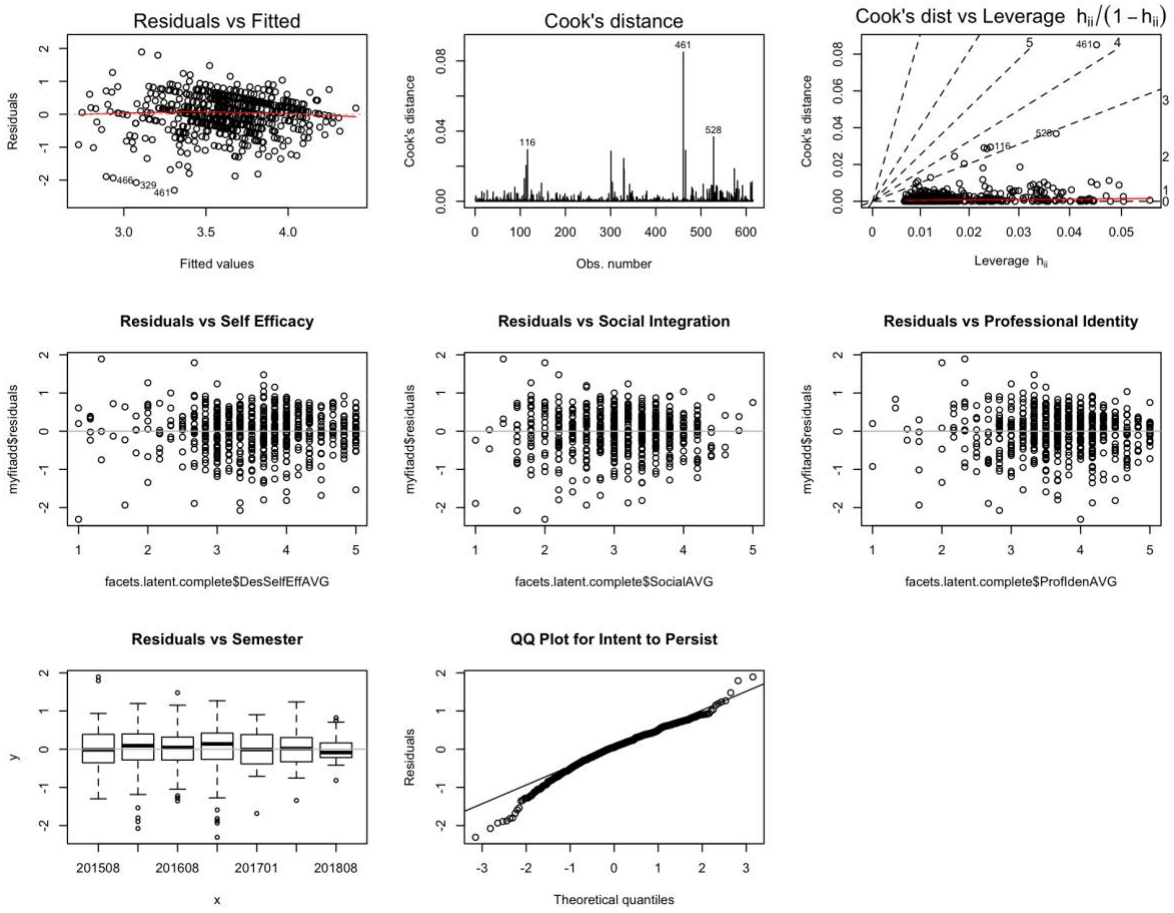

Figure 15. A set of graphs to check model assumptions in regression analysis. The discussion of the assumptions, which include normality, constant variance, and independence of the data, often involves looking for a random scatter of points about the dotted line, which we see in the residuals vs. fitted, residuals vs. each factor (self-efficacy, social integration, professional identity, and semester), and the residuals vs. order of data.

Checking the model assumptions of the regression model from the plots shown in Figure 15, we can summarize by stating the following:

- From the QQ plot, we can see that the plot looks fairly linear or maybe slightly S-shaped and is not scattered equally about the line. Therefore the QQplot shows a result that is not consistent with normality. The Shapiro-Wilk test's null hypothesis is that the population is normal. With the p-value (p-

value = 1.127×10^{-9}) significantly smaller than any alpha value (the smallest alpha value tends to be 0.01), the p-value shows that the normality assumption is violated.

- From the residuals vs. fitted plot, we can see the assumption of constant variance is possibly violated as the data points are clumped in a fairly elliptical shape. The Breusch-Pagan test (p-value = 4.04×10^{-11}), in which the null hypothesis is that the model has constant variance, confirms a violation of constant variance as well.
- Point 461 has a relatively large Cook's distance and seems to be a highly influential point (Cook's Distance vs. Leverage). Point 461 may therefore be an outlier. However, to exclude Point 461 would require further consultation with the data collector (Dr. Vanessa Svihla).

Essentially, the full additive model has some large problems, including violations of constant variance and normality.

Multicollinearity, a term that means there is redundancy among predictor variables and correlations exist between three or more predictor variables, is also a problem in the full model. Multicollinearity in a regression model can be detected using the variance inflation factor (VIF), which measures the inflation in the variance of the regression coefficients in the model due to multicollinearity. For the full additive model, the VIF values of all predictor variables was >1.3. As a general rule of thumb, if the VIF value is between 5 and 10, collinearity is a problem in the model. Therefore, multicollinearity is not a problem in the full additive model (James, Witten, Hastie, & Tibshirani, 2014).

The same regression analysis and ANOVA statistical methods were also applied to the shortened version of the dataset that exclusively had pre-test data, with the questions in Table 24 extracted. We will refer to this analysis as the short additive model. The response variable was still IntPers or the student's intent to persist within an engineering degree, and no interactions were incorporated within the model. The short regression table (Table 28) shows that none of the main effects contributed significantly to the short additive model except Semester.

Table 28. The statistical analysis of the linear regression is shown for the short additive model using data from the shortened survey for the pre-tests. The columns correspond to β_0(estimates), which correspond to the loadings on each variable or specific slope according to that variable; SE, or the standard error; t value, or the test run to determine the significance of each main effect to the overall model; and the p-value, the probability that we will obtain test results from the t-test that are at least as extreme as those observed.

Variable Name	β (Estimate)	SE	t value	p-value
Intercept (β_0)	3.015	0.259	11.660	$< 2 \times 10^{-16}$
DesSelfEffAVG	0.03210	0.0578	0.555	0.579
SocialAVG	-0.0271	0.0537	-0.504	0.615
ProfIdenAVG	0.0717	0.0633	1.133	0.259
Semester	0.189	0.0669	2.822	0.00525

The ANOVA statistics included in Table 29 confirm that none of main effects contribute significantly to the variation in the short additive model except for Semester. Therefore, we might infer from Tables 28 and 29 that further analysis is needed as we may need to transform variables within the linear regression, or we may need to revisit the use of linear regression as our model.

Table 29. ANOVA Type I Statistics. The alpha value for the factors in this dataset is set at 0.05 and any p-value that is below that alpha value contributes significantly to the variation in the model. Looking, then, at the p-

value column, we can see that all of the main effects contribute most significantly to the variation in the model.

ANOVA (Analysis of Variance)					
Source of Variation	**df**	**Sum of Squares (SS)**	**Mean Squares (MS = SS/df)**	**F value**	**p-value**
DesSelfEffAVG	1	0.399	0.399	1.779	0.184
SocialAVG	1	0.001	0.00141	0.0063	0.937
ProfIdenAVG	1	0.230	0.230	1.025	0.313
Semester	1	1.787	1.787	7.965	0.00525
Residuals	199	44.653	0.224		

This version of the data seems to have greatly improved the normality (Shapiro-Wilk p-value = 0.1916) and the constant variance (Breusch-Pagan p-value = 0.3568) of the model. The multicollinearity is still small (>1.4) between the main effects. However, as Semester is the only main effect that has significance in this model, the model is not a good fit for the data.

Conclusions

We began this analysis with several research questions that we intended to answer. Research question 1 asked whether Principal Component Analysis or Confirmatory Factor Analysis presents the most comprehensive data on which questions to eliminate from the survey. What we've realized in this analysis is that both methods present different ways of observing and analyzing patterns observed in the data. Both methods were instrumental in helping Dr. Svihla shorten the survey by ten questions. Research question 2 asked if the cluster analysis results confirmed the results of the PCA and CFA or if the results proposed different questions to eliminate from the survey. While the kmeans cluster analysis was not

as helpful in terms of eliminating potential survey questions, the *mclust* R package was far more helpful. However, the questions that *mclust* proposed to eliminate were totally different than the questions the CFA and PCA showed were most problematic. What statistical method was better then? The method (or set of methods) that the survey designer (and client) – Dr. Vanessa Svihla – found most informative and useful. In this case, those methods consisted of the CFA and PCA analyses. Research question 3 asked why these methods differed in the resulting analyses. This analysis has shown that the methods differed in their results because their methods and assumptions were different. CFA requires the clustering to be preset and for the data to obey certain assumptions (normality, independence, constant variance). PCA mixes the data to find linear combinations that account for the most variance possible. Cluster analysis uses machine learning to cluster and to see which data does not fit. While CFA and PCA are, to some degree, two sides of the same coin and therefore return similar results, cluster analysis is different and therefore returns different results. These underlying differences between the techniques help explain the differences in the results we obtained – that different methods proposed the elimination of different questions. Research question 4 asked if the shortened survey could be used to perform regression analysis and ANOVA techniques in an attempt to build a model to predict students' intent to persist in engineering. While I built an initial set of models, and the shortened survey certainly performed better in terms of meeting the assumptions, this model would need to be transformed to show significance among the main effects. Both linear models were problematic and, therefore, significant work is still needed to

build a model that both shows significance among the main effects (as the full additive model did) and does not violate the assumptions (as the short additive model did). Once the model is built, we'd need to see if the model accurately predicts future data collected.

Limitations and Future Work

The limitations of this study were numerous. While the data was analyzed using multiple techniques and the results mostly reiterated one another, some of the techniques used never yielded particularly helpful results. For instance, the PCA would need to be modified to perhaps lessen the number of variables included in the analysis. The kmeans cluster analysis never gave distinct clusters that were easy to evaluate. I used one-factor CFA for the latent variables, and perhaps that kind of analysis included too few items per factor. The regression analysis was admittedly limited in its scope in that I only evaluated a few averaged variables and I did not transform the data to look for a model that better fit the data. One might question whether linear regression was the right tool to model this data as well. Also, no matter how good the model for the data is, there is no statistical analysis that easily substitutes for the knowledge gained by regularly administering the survey, gathering informal participant feedback, and having a sense of which questions should be eliminated.

While quite a lot of data analysis was performed in this thesis, much work is still left. While much of the data was cleaned, the process was arduous and would need to be streamlined to clean the raw data still remaining and the raw data

regularly collected on a continuing basis. Potential ideas for future work include running a CFA, PCA, and cluster analysis on the new data collected from the shortened survey to compare the shortened survey data to the original long survey data and determine how the statistical analysis changed. We could also use theory to guide the removal of certain variables in addition to the statistical analyses. We could include the Spring data in the CFA, PCA, and cluster analyses so that our shortened survey includes models both semesters, not just one. Also, regression analysis could be greatly expanded upon using the latent variables originally used in this analysis or on different latent variables, including, perhaps, some of the principal components. Regression analysis could also use the shortened survey results to see if the model radically changes.

References

Adams, M. J. D., & Umbach, P. D. (2012). Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. *Research in Higher Education*, *53*(5), 576–591. https://doi.org/10.1007/s11162-011-9240-5

Bernold, L.E., Spurlin, J.E. and Anson, C.M. (2007). Understanding Our Students: A Longitudinal-Study of Success and Failure in Engineering with Implications for Increased Retention. *Journal of Engineering Education, 96*(3): 263-274. doi:10.1002/j.2168-9830.2007.tb00935.x

Brainard, S.G. & Carlin, L. (1998). A Six-Year Longitudinal Study of Undergraduate Women in Engineering and Science*. *Journal of Engineering Education, 87*(4): 369-375. doi:10.1002/j.2168-9830.1998.tb00367.x

Brems, M. (April 17, 2017). *A One-Stop Shop for Principal Component Analysis* [Blog Post]. Retrieved from https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Carberry, A.R., H.S. Lee, & M.W. Ohland. (2010). Measuring engineering design self-efficacy*. Journal of Engineering Education, 99*(1): 71-79.

Datye, A., Chi, E., Han, S., Svihla, V., & Kang, S. (2016). *NSF Award #1623105 – IUSE/PFE-RED: FACETS: Formation of Accomplished Chemical Engineers for Transforming Society* [Website]. Retrieved from https://www.nsf.gov/awardsearch/showAward?AWD_ID=1623105&HistoricalAwards=false

Everitt, B. S., & Dunn, G. (2001). *Applied Multivariate Data Analysis* (2nd edition). London: John Wiley & Sons, Ltd.

Forin, M. T. R., University, P., Lafayette, W., Adams, D. R., University, P., Lafayette, W., Hatten, K., & University, P. (n.d.). *Crystallized Identity: A Look at Identity Development through Cross-disciplinary Experiences in Engineering*. 22.

Fraley, C., Raftery, A. E., Murphy, B. T., & Scrucca, L. (2012). *Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. 597, 1-57. Retrieved from https://pdfs.semanticscholar.org/5bbc/022e371259d39cef9c47f453545a95cc36b2.pdf?_ga=2.86637596.501724156.1582476952-627374517.1582476952

Geisinger, B. N., & Raman, D. R. (2013). Why They Leave: Understanding Student Attrition from Engineering Majors*. International Journal of Engineering Education, 29*(4): 914-925.

Hartman, R. (March 22, 2017). *CFA in lavaan* [Blog post]. Retrieved from http://www.understandingdata.net/2017/03/22/cfa-in-lavaan/

Huang, G., Taddese, N., Walter, E., & Peng, S. (2000). *Entry and Persistence of Women and Minorities in College Science and Engineering Education* (NCES 2000–601). U.S. Department of Education. National Center for Education Statistics. Washington, DC.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: with applications in R.* Springer.

Johnson, R. A., & Wichern, D. W. (2015). *Applied Multivariate Statistical Analysis* (6th edition). Noida, India: Pearson India.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065): 20150202. https://doi.org/10.1098/rsta.2015.0202

Kenny, D. A. (Nov. 24, 2015). *Measuring Model Fit* [Website]. Retrieved from http://davidakenny.net/cm/fit.htm#null

Kodali, T. (2015). *K Means Clustering in R: R-Bloggers* [Blog Post]. Retrieved from https://www.r-bloggers.com/k-means-clustering-in-r/

Mosborg, S., R.S. Adams, R. Kim, C.J. Atman, J. Turns, and M.E. Cardella. (2005). *Conceptions of the Engineering Design Process: An Expert Study of Advanced Practicing Professionals*, in *Proceedings of ASEE Annual Conference & Exposition*. ASEE: Portland, OR. 1-27.

Nocito-Gobel, J., M.A. Collura, S. Daniels, and I. Orabi. (2005). *Are Attitudes Toward Engineering Influenced by a Project-Based Introductory Course*, in *Proceedings of ASEE Annual Conference and Exposition: The Changing Landscape of Engineering and Technology Education in a Global World*. ASEE: Portland, OR. 693-706.

Pierrakos, O., T.K. Beam, J. Constantz, A. Johri, and R. Anderson. (2009). *On the development of a professional identity: engineering persisters vs engineering switchers*. in *Proceedings of the 39th Frontiers in Education Conference*. IEEE: San Antonio, TX.

Porter, S. R. (2004). Raising response rates: What works? *New Directions for Institutional Research*, *2004*(121), 5–21. https://doi.org/10.1002/ir.97

Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, *2004*(121), 63–73. https://doi.org/10.1002/ir.101

Santiago, D. L. Y., & Hensel, R. (2012). *Engineering Attrition and University Retention*. ASEE: San Antonio, TX. 693-706.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal*, *8*(1), 289–317. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/

Sheppard, S., Gilmartin, S., H.L. Chen, K. Donaldson, G. Lichtenstein, O. Eris, M. Lande, & Toye, G. (2010). *Exploring the Engineering Student Experience: Findings from the Academic Pathways of People Learning Engineering Survey (APPLES). TR-10-01.* Center for the Advancement of Engineering Education (NJ1).

Sinickas, A. (2007). Finding a cure for survey fatigue. *Strategic Communication Results*, *11*(2), 11. https://search.proquest.com/openview/2d5769d2c0348b8238763c1b46debe2c/1?pq-origsite=gscholar&cbl=44514

Study Design. (n.d.). FACETS: Formation of Accomplished Chemical Engineers for Transforming Society [Website]. Retrieved April 7, 2020, from http://facets.unm.edu/research/study-design.html

Tracy, S. J., & Trethewey, A. (2005). Fracturing the Real-Self<-->Fake-Self Dichotomy: Moving Toward “Crystallized” Organizational Discourses and Identities. *Communication Theory*, *15*(2), 168–195. https://doi.org/10.1111/j.1468-2885.2005.tb00331.x

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/